# Contents

1 Empirical risk minimization .............................................. 5  
1.1 Setting ......................................................................... 5  
1.1.1 Supervised learning .................................................. 6  
1.1.2 Unsupervised learning ............................................... 7  
1.2 Bayes estimator ............................................................ 10  
1.3 Empirical risk minimization ............................................ 12  
1.4 Hoeffding’s inequality .................................................... 13  
1.5 Learning from finite dictionaries ..................................... 15  
1.6 Benett’s and Bernstein’s inequalities .................................. 17  
1.7 Fast rates ..................................................................... 19  
1.7.1 Fast rates under Bernstein’s condition ......................... 19  
1.7.2 Examples satisfying the Bernstein’s assumption ............ 20  

2 Learning with infinite ressources ....................................... 23  
2.1 The general framework .................................................. 23  
2.1.1 Example: Lipschitz loss and linear functionals ............... 23  
2.2 Azuma-Hoeffding’s inequality ......................................... 24  
2.3 The bounded difference inequality ................................... 25  
2.4 Symmetrization principle ............................................... 25  
2.5 Application to learning theory ........................................ 26  
2.6 The finite case ................................................................ 27  
2.7 Vapnik-Chervonenkis theory ............................................ 28  
2.7.1 VC inequality ........................................................... 28  
2.8 Covering numbers ........................................................ 29  
2.9 The chaining method ..................................................... 32  

3 Convex relaxation ............................................................. 35  
3.1 Construction of convex proxy ......................................... 35  
3.2 Link with hard classifiers ............................................... 37  
3.3 Bounding the $\varphi$-excess risk ....................................... 41  
3.4 Boosting ...................................................................... 42
## CONTENTS

### 3.5 Support Vector Machine
- 3.5.1 Reproducing kernel Hilbert space
- 3.5.2 Representer theorem
- 3.5.3 Excess risk of $\varphi$-ERM

### 4 Convex optimization
- 4.1 Convex problems
- 4.2 Gradient descent
- 4.3 Projected gradient descent
- 4.4 Examples
  - 4.4.1 SVM
  - 4.4.2 Boosting
- 4.5 Mirror descent
- 4.6 Stochastic optimization
  - 4.6.1 Stochastic gradient descent
  - 4.6.2 Stochastic mirror descent
Chapter 1

Empirical risk minimization

1.1 Setting

Let $Z$ denote a measurable space where data take values. Let $F$ denote a set of parameters. For any $f \in F$ and $z \in Z$, let $\ell_f(z)$ denote the loss of parameter $f$ at observation point $z$. Let $P$ denote a probability distribution on $Z$. The risk of any parameter $f \in F$ is measured by

$$R(f) = P\ell_f := \mathbb{E}_{Z \sim P}[\ell_f(Z)].$$

This course deals with minimizing the risk function

$$\min_{f \in F} P\ell_f.$$

The distribution $P$ is unknown, so exact minimization of the risk is impossible. Nevertheless, minimizers of the risk will play an important role. When it exists, $f^*_P \in \arg\min_{f \in F} P\ell_f$ is called an oracle. To approximate $f^*_P$, a data set $\mathcal{D}_N = \{Z_1, \ldots, Z_N\}$ is available. In these notes, $\mathcal{D}_N$ always contains i.i.d. random variables with distribution $P$. A natural idea to estimate $f^*_P$ is to estimate each loss $P\ell_f$ by its empirical estimator $P_N\ell_f = N^{-1} \sum_{i=1}^N \ell_f(Z_i)$. Then, $f^*_P$ is estimated by the empirical risk minimizer

$$\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N\ell_f.$$

If a minimizer does not exist, one can define alternatively, for $\epsilon > 0$, $\epsilon$-minimizers $\hat{f}^{(\epsilon)}_{\text{erm}}$ as any function satisfying

$$P_N\ell_{\hat{f}^{(\epsilon)}_{\text{erm}}} \leq \inf_{f \in F} P_N\ell_f + \epsilon.$$

Most of these notes deal with these estimators. Empirical risk minimization is quite common as can be appreciated from the following examples.
CHAPTER 1. EMPIRICAL RISK MINIMIZATION

1.1.1 Supervised learning

In supervised learning, data \( z \in Z \) are couples \( z = (x, y) \in X \times Y \) and parameters \( f \in F \) are functions \( f : X \to Y \). The most classical problems in supervised learning are classification where \( Y \) is a discrete set as in binary classification where \( Y = \{0, 1\} \) and regression where \( Y \subset \mathbb{R} \) is a continuous subset.

Among classical loss functions, one can mention the 0-1 loss in classification \( \ell_f(x, y) = 1_{y \neq f(x)} \) and the quadratic loss in regression \( \ell_f(x, y) = (y - f(x))^2 \).

In the couple \((x, y)\), \( x \) is called the input and \( y \) the output. The interpretation of supervised learning is that one wants to predict a typical output \( Y \) associated to the input \( X \) when \((X, Y) \sim P\).

For example, classification algorithms are routinely used to classify spams or help diagnosis.

In supervised learning, the function minimizing the risk \( P\ell_f \) among all functions \( f : X \to Y \) for which this risk is well defined would be an ideal estimator. It is referred to as the Bayes estimator and it is denoted by \( f^* \).

For any \( f \in F \), one has the relations

\[
R(f^*) \leq R(f^*_F) \leq R(f) .
\]

The difference \( \mathcal{E}(f) := R(f) - R(f^*) \) is usually called the excess risk of \( f \). Let us also introduce the notation \( \mathcal{E}_{\ell, F}(f) = R(f) - R(f^*_F) \) which is non negative for any \( f \in F \), so the excess risk can be decomposed

\[
\mathcal{E}(f) = \mathcal{E}(f^*_F) + \mathcal{E}_{\ell, F}(f) .
\]

In particular, for any data driven \( \hat{f} \in F \),

\[
\mathcal{E}(\hat{f}) = \mathcal{E}(f^*_F) + \mathcal{E}_{\ell, F}(\hat{f}) .
\]

The term \( \mathcal{E}(f^*_F) \) is an error unavoidable, sometimes called bayes of estimation. Implicitly, working with the set of functions \( F \) means that this error is assumed to be small. Of course, the richest the class \( F \), the larger is the class \( f^* \) of functions for which it is true. The other term \( \mathcal{E}_{\ell, F}(\hat{f}) = R(\hat{f}) - R(f^*_F) \) is random as \( R(\hat{f}) \) is (the expectation defining \( R(f) \) is only taken with respect to \( Z \), independent of \( D \)), so

\[
R(\hat{f}) = \mathbb{E}_{Z \sim P}[\hat{f}(Z)|D_N] .
\]

A large part of these notes aims at providing deterministic upper bounds on \( \mathcal{E}_{\ell, F}(f^*_\text{erm}), \Delta_{N, \delta}(F) \) such that, with probability at least \( 1 - \delta \),

\[
\mathcal{E}_{\ell, F}(f^*_\text{erm}) \leq \Delta_{N, \delta}(F) .
\]
One will also sometimes try to bound the expectation $\mathbb{E}[\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}})] \leq \Delta_N(F)$. These inequalities are referred to as *oracle inequalities* as they compare the risk of the estimator $\hat{f}_{\text{erm}}$ with the one of the oracle $f_F^*$. The term $\Delta_{N,\delta}(F)$ typically goes to 0 when $N \to \infty$, while it grows with the complexity of $F$.

It is tempting to minimize the excess risk to choose a small set $F$ (the excess risk is null if $F$ is a singleton), but, even if $\hat{f}_{\text{erm}}$ has a small excess risk, it may have poor prediction properties if the class $F$ is not rich enough. Ideally, the model $F$ should be chosen to realize a tradeoff between the errors $\mathcal{E}(f_F^*)$ and $\mathcal{E}_{\ell,F}(\hat{f})$. These lectures aim at understanding the error $\mathcal{E}_{\ell,F}(\hat{f})$ in this tradeoff.

### 1.1.2 Unsupervised learning

In unsupervised learning, we don’t observe the output $y$ so data are simply inputs $z = x \in \mathcal{X}$. One wants to learn features of their distribution $P$.

**Multivariate mean estimation** Let $\mathcal{X} \subset \mathbb{R}^d$, and imagine that one wants to estimate the expectation $f_F^* = P[Z]$. Let then $F = \mathbb{R}^d$ and $\ell_f(x) = \|x - f\|^2$, where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$. For any $f \in F$

$$
\|X - f\|^2 = \|X - f_F^*\|^2 + \|f - f_F^*\|^2 - 2(X - f_F^*)^T(f - f_F^*) .
$$

Hence

$$
P \ell_f = P[\|X - f\|^2] = P[\|X - f_F^*\|^2] + \|f - f_F^*\|^2 \geq P[\|X - f_F^*\|^2] = P \ell_{f_F^*} .
$$

Thus

$$
\arg\min_{f \in F} P \ell_f = \{f_F^*\}, \quad \mathcal{E}_{\ell,F}(f) = \|f - f_F^*\|^2 .
$$

In this example one can check that $\hat{f}_{\text{erm}}$ is simply the empirical mean

$$
\hat{f}_{\text{erm}} = \frac{1}{N} \sum_{i=1}^N X_i .
$$

**Least-squares density estimation** Assume that $P$ has density $f^*$ w.r.t. a known measure $\mu$ on $\mathcal{X}$ and that $f^* \in L^2(\mu)$. Let $(\varphi_i)_{i=1,\ldots,d}$ denote an orthonormal system in $L^2(\mu)$ and let $F$ denote the linear span of the $(\varphi_i)_{i=1,\ldots,d}$. Let $f_F^*$ denote the orthogonal projection of $f^*$ onto $F$ and let

$$
\forall f \in L^2(\mu), \quad \ell_f(x) = \|f\|_{L^2(\mu)}^2 - 2f(x) .
$$
The key is to remark that, for any \( f \in L^2(\mu) \),

\[
\langle f^*, f \rangle_{L^2(\mu)} = \int f f^* \, d\mu = P[f],
\]

and, for any \( f \in F \), by definition of \( f_F^* \),

\[
\langle f^*, f \rangle_{L^2(\mu)} = \langle f_F^*, f \rangle_{L^2(\mu)}.
\]

From these relationships,

\[
P\ell = \|f\|_{L^2(\mu)}^2 - 2Pf = \|f\|_{L^2(\mu)}^2 - 2 \langle f^*, f \rangle_{L^2(\mu)}
\]

\[
= \|f\|_{L^2(\mu)}^2 - 2 \langle f_F^*, f \rangle_{L^2(\mu)} = \|f - f_F^*\|_{L^2(\mu)}^2 - \|f_F^*\|_{L^2(\mu)}^2
\]

\[
\geq - \|f_F^*\|_{L^2(\mu)}^2 = P\ell_{f_F^*}.
\]

Hence, \( f_F^* \) is the oracle in \( F \) (unique \( \mu \)-a.s.). One can also check that

\[
f_F^* = \sum_{i=1}^d P[\varphi_i] \varphi_i, \quad \hat{f}_{\text{erm}} = \sum_{i=1}^d P[\varphi_i] \varphi_i.
\]

The excess risk is \( E(f) = \|f - f^*\|_{L^2(\mu)}^2 \) and

\[
E_{\ell,F}(\hat{f}_{\text{erm}}) = \|\hat{f}_{\text{erm}} - f_F^*\|_{L^2(\mu)}^2 = \sum_{i=1}^d ((P_N - P) \varphi_i)^2.
\]

This problem has strong connections with multivariate means estimation. Since

\[
f_F^* = \sum_{i=1}^d P[\varphi_i] \varphi_i,
\]

estimating \( f_F^* \) amounts to estimate the finite dimensional expectation

\[
\mathbf{f}_F = P \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_d \end{bmatrix} \in \mathbb{R}^d.
\]

The estimator provided in the last section was

\[
\hat{f}_{\text{erm}} = P_N \begin{bmatrix} \varphi_1 \\ \vdots \\ \varphi_d \end{bmatrix} \in \mathbb{R}^d.
\]
1.1. SETTING

The natural embedding of this estimator is thus

$$\hat{f}_{\text{erm}} = \sum_{i=1}^{d} P_N[\varphi_i] \varphi_i.$$  

Moreover, (1.1) shows that the excess risk of this estimator is the squared Euclidean distance between $f^*_F$ and $\hat{f}_{\text{erm}}$. There are important differences between both problems though. First, in least-squares density estimation, any estimator $\hat{f} \in F$ usually has a bias $\|f^* - f^*_F\|^2$ while this term in null in multivariate mean estimation. Second, in density estimation, the statistician has the choice of the basis $\varphi_i$ and the dimension $d$. The functions $\varphi_i$ can therefore be chosen bounded, which implies that the random variables $\varphi_i(X_j)$ are bounded. On the other for multivariate mean estimation, the coordinates $X_{i,j}$ can be unbounded.

**Maximum likelihood estimators** To estimate $P$, an alternative to the projection method, widely spread in statistics, is to start with a set $F$ of densities with respect to a given measure $\mu$ (like Lebesgue measure on $\mathbb{R}^d$) and define, for each $f$, the likelihood of the data set $D_N$:

$$L_f(D_N) = \prod_{i=1}^{N} f(X_i).$$

The Maximum Likelihood Estimator is a maximizer of the likelihood

$$\hat{f}_{\text{erm}} \in \arg\max_{f \in F} L_f(D_N).$$

Equivalently, $\hat{f}_{\text{erm}}$ is the minimizer of

$$\frac{1}{N} (- \log(L_f(D_N))) = P_N[- \log(f(X))].$$

Define $\ell_f(x) = - \log(f(x))$ so $\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f$ is an empirical risk minimizer. The risk $P\ell_f$ associated to this loss is therefore

$$P\ell_f = P[- \log(f(X))].$$

To check that is is a natural risk measure, one can compare its value at some $f \in F$ and in the true density $f^*$. It holds

$$P[\ell_f - \ell_{f^*}] = \int f^* \log \left( \frac{f}{f^*} \right) d\mu.$$

It is the Kullback divergence between $f^*$ and $f$, it is always non-negative as, by Jensen’s inequality, since log is concave,

$$\int f^* \log \left( \frac{f}{f^*} \right) d\mu = \mathbb{E} \left[ \log \left( \frac{f(X)}{f^*(X)} \right) \right] \leq \log \left( \mathbb{E} \left[ \frac{f(X)}{f^*(Z)} \right] \right) = \log \left( \int f d\mu \right) = 0.$$
CHAPTER 1. EMPIRICAL RISK MINIMIZATION

Clustering A connection with classification can be made when the distribution $P$ of $X$ is a mixture. In the simplest case, say that there exist $p \in (0, 1)$, $f_0$ and $f_1$ such that the density $f$ of $X$ w.r.t. a given measure $\mu$ on $\mathcal{X}$ is

$$f = (1 - p)f_0 + pf_1.$$ 

One way to realize this distribution is to say that $X$ has the distribution of the first marginal of the couple $(X, Y)$, whose distribution is defined as follows: $Y$ has Bernoulli distribution $Y \sim \mathcal{B}(p)$ and, conditioning on $Y = 1$, $X \sim f_1$ while, conditioned on $Y = 0$, $X \sim f_0$. The clustering problem is to find the distribution of $Y|X$: given the observation $x$, one wants to know if this random variable $X$ that generated this observation has been simulated using $f_0$ (in this case, say that $X$ belongs to the class 0) or $f_1$ (in this case, $X$ belongs to the class 1). This problem looks like binary classification, except that the outputs $Y_i$ associated to the inputs $X_i$ are not observed and cannot be used to train the estimators. Clustering is a classification problem without observation (or supervision) of the outputs. This explains the name “unsupervised” learning.

1.2 Bayes estimator

In the remaining of this chapter, we focus on binary classification, which is a supervised learning problem where the output space $\mathcal{Y} = \{0, 1\}$. The set of parameters $F$ is a set of classifiers, that is functions $f : \mathcal{X} \to \{0, 1\}$. The loss $\ell_f$ is the 0–1 loss $\ell_f(x, y) = 1_{(f(x) \neq y)}$.

Let us first assume that $F$ is the set of all functions $f : \mathcal{X} \to \mathcal{Y}$ and call $\eta$ the regression function, that is a function such that, for any measurable bounded function $\varphi : \mathcal{X} \to \mathbb{R}$, $\mathbb{E}[\varphi(X)] = \mathbb{E}[\eta(X)\varphi(X)]$.

**Definition 1.** The Bayes classifier $f^*$ is defined by

$$\forall x \in \mathcal{X}, \quad f^*(x) = 1_{\{\eta(x) > 1/2\}}.$$ 

The Bayes classifier minimizes the classification error among all functions $f : \mathcal{X} \to \mathcal{Y}$ as shown by the following theorem.

**Theorem 2.** The risk of the Bayes classifier satisfies

$$R(f^*) = \mathbb{E}[\min(\eta(X), (1 - \eta(X)))] \leq 1/2.$$ 

Moreover, for any $f : \mathcal{X} \to \{0, 1\}$, the excess risk of $f$ defined as $\mathcal{E}(f) := R(f) - R(f^*)$ satisfies

$$\mathcal{E}(f) = \mathbb{P}[2\eta - 1|1_{(f \neq f^*)}] = \mathbb{E}[2\eta(X) - 1|1_{(f(X) \neq f^*(X))}].$$
1.2. BAYES ESTIMATOR

Proof. By definition, for any function $f: \mathcal{X} \to \mathcal{Y}$,

$$R(f) = \mathbb{E}[Y(1-f(X)) + (1-Y)f(X)]$$

$$= \mathbb{E}[\eta(X)(1-f(X)) + (1-\eta(X))f(X)] .$$

By definition of the Bayes classifier, it follows that

$$R(f^*) = \mathbb{E}[\eta(X)1_{\{\eta(X)\leq 1/2\}} + (1-\eta(X))1_{\{1-\eta(X)\leq 1/2\}}]$$

$$= \mathbb{E}[\min(\eta(X), (1-\eta(X)))] .$$

As $\min(\eta(X), (1-\eta(X))) \leq 1/2$ almost surely, $R(f^*) \leq 1/2$ and $R(f^*) = 1/2$ iff $\eta(X) = 1/2$ almost surely, which happens when $X$ does not bring information of $Y$.

Moreover, for any function $f: \mathcal{X} \to \mathcal{Y}$,

$$R(f) - R(f^*) = \mathbb{E}[\eta(X)(f^*(X) - f(X)) + (1-\eta(X))(f(X) - f^*(X))]$$

$$= \mathbb{E}[(2\eta(X) - 1)(f^*(X) - f(X))] .$$

Now, $(f^*(x) - f(x)) = 0$ if $f^*(x) = f(x)$, $f^*(x) - f(x) = 1$ if $\eta(x) > 1/2$ and $f^*(x) \neq f(x)$ and $f^*(x) - f(x) = -1$ if $\eta(x) \leq 1/2$ and $f^*(x) \neq f(x)$. Overall, $f^*(x) - f(x) = 1_{\{f^*(x)\neq f(x)\}}\text{sign}(2\eta(x) - 1)$ and

$$(2\eta(X) - 1)(f^*(X) - f(X)) = |2\eta(X) - 1|1_{\{f^*(x)\neq f(x)\}} .$$

Theorem 2 implies that $f^*$ is the best possible classifier. However, this classifier depends on the unknown distribution $P$ of the data and is not available in practice. Instead, we want to build a random function $\hat{f}: \mathcal{X} \to \mathcal{Y}$ such that the excess risk $\mathcal{E}(\hat{f}) = R(\hat{f}) - R(f^*)$ is small. This task is hard in general as shown by the following negative result.

Theorem 3. If $\mathcal{X}$ is infinite, then, for any $N \geq 1$, for any classifier $\hat{f}_N$ built with $\mathcal{D}_N$, for any $\epsilon > 0$, there exists a distribution $P_N$ on $\mathcal{X} \times \mathcal{Y}$ such that $R(f^*) = 0$ and $\mathcal{E}(\hat{f}_N) \geq 1/2 - \epsilon$.

The proof is left as an exercise. One can for example proceed by completing the following steps. Let $n$, $K$ denote integers, $f: (\mathcal{X} \times \mathcal{Y})^n \mapsto \{0, 1\}$ a classifier. The set $\mathcal{X}$ being infinite, one can pick $K$ distinct points $a_1, \ldots, a_K$ in $\mathcal{X}$. Denote by $(\mathbb{P}_q)_{q \in \mathcal{Y}^K}$ denote the set of probability distributions on $\mathcal{X} \otimes \mathcal{Y}$, such that, for any $q \in \mathcal{Y}^K$,

$$\mathbb{P}_q(X = a_j, Y = q_j) = K^{-1} \quad \text{for any } j \in \{1, \ldots, K\}.$$
1. Show that, for any \( q \in \{0, 1\}^N \), \( \mathbb{P}_q(X = a_j) = 1/K \) for any \( j \in \{1, \ldots, K\} \).

2. Compute \( \mathbb{P}_q(Y \neq f(X)) \) for any classifier \( f \).

3. Show that \( L_q^* = \inf_{f: X \to Y} \mathbb{P}_q(Y \neq f(X)) = 0 \) and that the infimum is achieved by any \( f_q^* \), such that, for any \( x \in \{a_1, \ldots, a_K\} \),

\[
f_q^*(x) = \sum_{i=1}^{K} q_i 1_{a_i}(x).
\]

4. Show that

\[
\sup_{\mathbb{P}} \{ \mathbb{E}_{\mathbb{P}^n} [\mathbb{P}(Y \neq f(X; D_n)|D_n)] - L_q^* \} \geq \sup_{q \in \mathbb{Y}^K} \mathbb{E}_{\mathbb{P}^n} [\mathbb{P}_q(Y \neq f(X; D_n)|D_n)]
\]

\[
\geq \frac{1}{2K} \sum_{q \in \mathbb{Y}^K} \mathbb{E}_{\mathbb{P}^n} [\mathbb{P}_q(Y \neq f(X; D_n)|D_n)]
\]

\[
\geq \frac{1}{2K} \sum_{q \in \mathbb{Y}^K} \mathbb{E}_{\mathbb{P}^n} [\mathbb{E}_{\mathbb{P}_q} [1_{Y \neq f(X; D_n)|X, D_n} 1_{X \neq \{X_1, \ldots, X_n\}}]]
\]

5. Show that, for any \( x \notin \{X_1, \ldots, X_n\} \), there exist \( 2^{K-1} \) values of \( q \in \{0, 1\}^K \) such that \( \mathbb{E}_{\mathbb{P}_q} [1_{Y \neq f(x; D_n)|D_n}] = 0 \) and \( 2^{K-1} \) values of \( q \in \{0, 1\}^K \) such that \( \mathbb{E}_{\mathbb{P}_q} [1_{Y \neq f(x; D_n)|D_n}] = 1 \).

6. Show that, for any \( q \in \{0, 1\}^K \) and for any \( x \in \{a_1, \ldots, a_K\} \),

\[
\mathbb{E}_{\mathbb{P}_q} [1_{X_1 \neq x, X_2 \neq x, \ldots, X_n \neq x}] = (1 - 1/K)^n.
\]

7. Conclude.

The important message carried by Theorem 3 is that one cannot hope to bound \( \mathcal{E}(\hat{f}) \) without making assumption on \( P \). In the following, these assumptions will be made via a choice of a subset \( F \). We will implicitly assume that \( \mathcal{E}(f_F^*) \) is not too large and focus on bounding \( \mathcal{E}_{\ell,F}(\hat{f}) \).

### 1.3 Empirical risk minimization

Hereafter, we assume that a subset \( F \) of functions \( f : \mathcal{X} \to \mathcal{Y} \) is given, and, assuming that such a function exists, we denote by \( f_F^* \) an “oracle”: \( f_F^* \in F \) and satisfies

\[
f_F^* \in \operatorname{argmin}_{f \in F} R(f).
\]
The oracle cannot be used as a predictor either, but one can hope to estimate it correctly if $F$ is not too large. Recall that the excess risk of any estimator $\hat{f} \in F$ is decomposed as follows

$$\mathcal{E}(\hat{f}) = \mathcal{E}(f^*_F) + \mathcal{E}_{\ell,F}(\hat{f}).$$

(1.2)

The error $\mathcal{E}_{\ell,F}(\hat{f}) = R(\hat{f}) - R(f^*_F)$ is a stochastic term called estimation error that one hopes to bound either in expectation or by a deterministic quantity $\Delta_{N,\delta}(F)$ with probability at least $1 - \delta$. The residual excess risk $\mathcal{E}(f^*_F)$ is a modelisation error unavoidable for any $\hat{f} \in F$ that is comparable to the bias in statistics.

We focus on a particular estimator called empirical risk minimizer, that was originally introduced by Vapnik. The idea is simple, the oracle $f^*_F$ minimizes $P\ell_f$ over $F$. The operator $P$ is unknown in this definition but can be estimated by the empirical mean operator $P_N$ defined for any real valued function $\varphi : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$ by $P_N \varphi = (1/N) \sum_{i=1}^N \varphi(X_i, Y_i)$. The unknown risk $P\ell_f$ can be estimated by the empirical risk $P_N \ell_f$ and $f^*_F$ by a minimizer $\hat{f}_{\text{erm}}$ of the empirical risks

$$\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f, \quad P_N \ell_f = \frac{1}{N} \sum_{i=1}^N 1_{(f(X_i) \neq Y_i)}.$$

When exact minimization is hard, one can also define $\epsilon$-ERM $\hat{f}_{\text{erm}}(\epsilon)$ such that

$$P_N \ell_{\hat{f}_{\text{erm}}(\epsilon)} \leq \min_{f \in F} P_N \ell_f + \epsilon.$$

### 1.4. Hoeffding’s Inequality

To bound the risk of $\hat{f}_{\text{erm}}$, it is important to understand the deviation of the empirical mean around it’s expectation. Hoeffding’s inequality is the first among several inequalities that achieves such a job.

**Theorem 4** (Hoeffding’s inequality). Assume that $Z_1, \ldots, Z_N$ are independent random variables and let $f_1, \ldots, f_N$ denote real valued functions, $f_i$ taking values in $[a_i, b_i]$. Then

$$\forall t > 0, \quad \mathbb{P}\left(\frac{1}{N} \sum_{i=1}^N (f_i(Z_i) - \mathbb{E}[f_i(Z_i)]) > t\right) \leq e^{-2Nt^2/\sum_{i=1}^N (b_i - a_i)^2}.$$

**Proof.** Remark that one can assume without loss of generality that $\mathbb{E}[f_i(Z_i)] = 0$. The proof follows classical steps that we will stress.
CHAPTER 1. EMPIRICAL RISK MINIMIZATION

Lemma 5 (Chernoff bound). Let \( U \) denote a real valued random variable such that \( \mathbb{E}[e^{sU}] < \infty \) for any \( s \in (u, v) \subset \mathbb{R}_+ \), then

\[
\forall t > 0, \quad \mathbb{P}(U > t) \leq e^{-\psi_U^*(t)},
\]

where the Fenchel-Legendre transform of \( U \) is defined by

\[
\psi_U^*(t) = \sup_{s \in (u,v)} st - \log \mathbb{E}[e^{sU}].
\]

Proof. As the functions \( x \mapsto e^{sx} \) are nondecreasing,

\[
\forall s \in (u,v), \ t > 0, \quad \mathbb{P}(U > t) = \mathbb{P}(e^{sU} > e^{st}).
\]

By Markov's inequality,

\[
\forall s \in (u,v), \ t > 0, \quad \mathbb{P}(U > t) = e^{-st + \log \mathbb{E}[e^{sU}]}.
\]

The result follows by optimizing over \( s \in (u,v) \).

By the Chernoff bound, for any \( t > 0 \),

\[
\mathbb{P}\left( \frac{1}{N} \sum_{i=1}^{N} (f_i(Z_i) - \mathbb{E}[f_i(Z_i)]) > t \right) \leq e^{-\sup_{s > 0} st - \log \mathbb{E}[e^{s\sum_{i=1}^{N} f_i(Z_i)}]}.
\]

By independence of the \( Z_i \), it follows that the Laplace transform

\[
\log \mathbb{E}[e^{s/\mathbb{E}[\sum_{i=1}^{N} f_i(Z_i)]}] = \sum_{i=1}^{N} \log \mathbb{E}[e^{s f_i(Z_i) / N}] \quad (1.3)
\]

Therefore, the result will be obtained by bounding from above the Laplace transform of bounded variable. This is the purpose of the following lemma.

Lemma 6 (Hoeffding’s Lemma). Let \( U \) denote a random variable taking values in \([a, b]\) with distribution \( \mathbb{E} \). Then

\[
\forall s > 0, \quad \log \mathbb{E}[e^{sU}] \leq \frac{s^2(b-a)^2}{8}.
\]

Proof. The Laplace transform of \( U \), \( \psi_U(s) = \log \mathbb{E}[e^{sU}] \) is defined for any \( s > 0 \), it is a convex function infinitely differentiable such that \( \psi_U(0) = \psi_U'(0) = 0 \). Moreover, for any \( s > 0 \),

\[
\psi_U'(s) = \mathbb{E}\left[ U e^{sU} \frac{1}{\mathbb{E}[e^{sU}]} \right],
\]

\[
\psi_U''(s) = \mathbb{E}\left[ U^2 e^{sU} \frac{1}{\mathbb{E}[e^{sU}]} \right] - \left( \mathbb{E}\left[ U e^{sU} \frac{1}{\mathbb{E}[e^{sU}]} \right] \right)^2.
\]
As $e^{sU}/\mathbb{E}[e^{sU}]$ is a nonnegative random variable such that $\mathbb{E}[e^{sU}/\mathbb{E}[e^{sU}]] = 1$, one can consider the distribution $\mathbb{F}$ such that $d\mathbb{F} = (e^{sU}/\mathbb{E}[e^{sU}])d\mathbb{E}$. Therefore, $\psi''_U(s) = \text{Var}_\mathbb{F}(U) = \text{Var}_\mathbb{F}(U - (b - a)/2)$. As the random variable $|U - (b - a)/2| \leq (b - a)/2$ almost surely, we have

$$\text{Var}_\mathbb{F}(U - (b - a)/2) \leq \frac{(b - a)^2}{4}.$$ 

Applying the fundamental theorem of analysis, it follows

$$\psi_U(s) = \psi_U(s) - \psi_U(0) = \int_0^s \psi'_U(t)dt$$

$$= \int_0^s (\psi'_U(t) - \psi'_U(0))dt = \int_0^s \int_0^t \psi''_U(u)dudt$$

$$\leq \int_0^s \int_0^t \frac{(b - a)^2}{4}dudt = \frac{\int_0^s (b - a)^2}{4}tdt = \frac{(b - a)^2s^2}{8}.$$ 

Plugging the result of Hoeffding’s Lemma in Eq (1.3) yields

$$\log \mathbb{E}[e^{(s/N)\sum_{i=1}^N f_i(Z_i)}] = \frac{s^2}{2} \sum_{i=1}^N \frac{(b_i - a_i)^2}{4N^2}.$$ 

To conclude we use the simple following result

$$\forall a \in \mathbb{R}, b > 0, \quad \sup \left\{ ax - \frac{bx^2}{2} \right\} = -\frac{a^2}{2b}. \quad (1.5)$$

Together with (1.4), (1.5) shows that the Fenchel-Legendre transform of $U = 1/N \sum_{i=1}^N f_i(Z_i)$ is bounded from bellow by

$$\psi^*_U(t) \geq \frac{t^2}{2} \sum_{i=1}^N \frac{(b_i - a_i)^2}{4N^2} = \frac{2N^2t^2}{\sum_{i=1}^N (b_i - a_i)^2}.$$ 

The theorem follows from the Chernoff bound.

### 1.5 Learning from finite dictionaries

This section shows a first oracle inequality for the empirical risk minimizer $\hat{f}_{\text{erm}}$ in the elementary case where $F = \{f_1, \ldots, f_M\}$. 


Theorem 7. For any \( \epsilon > 0 \), \( \epsilon \)-ERM over \( F = \{ f_1, \ldots, f_M \} \) satisfy,

\[
\mathbb{E}[\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}})] \leq \sqrt{\frac{2\log(2M)}{N}} + \epsilon ,
\]

and, for all \( \delta \in (0, 1) \)

\[
\mathbb{P}
\left(
\mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}) > \sqrt{\frac{2\log(2M/\delta)}{N}} + \epsilon
\right) \leq \delta .
\]

Proof. The proof relies on the elementary but useful result of Vapnik

Lemma 8 (Vapnik’s Lemma). For any \( \epsilon > 0 \), \( \epsilon \)-ERM satisfy, almost surely,

\[
\mathbb{E}[\mathcal{E}_{\ell,F}(\hat{f}^{(\epsilon)}_{\text{erm}})] \leq 2 \sup_{f \in F} |(P_N - P)\ell_f| + \epsilon .
\]

Proof. By definition of \( \hat{f}^{(\epsilon)}_{\text{erm}} \), \( P_N[\ell^{(\epsilon)}_{\text{erm}} - \ell_f^*] \leq \epsilon \), therefore,

\[
\mathbb{E}[\mathcal{E}_{\ell,F}(\hat{f}^{(\epsilon)}_{\text{erm}})] = (P - P_N)\ell^{(\epsilon)}_{\text{erm}} + P_N[\ell^{(\epsilon)}_{\text{erm}} - \ell_f^*] + (P_N - P)\ell_f^* \\
\leq (P_N - P)[\ell_f^* - \ell^{(\epsilon)}_{\text{erm}}] + \epsilon \\
\leq 2 \sup_{f \in F} |(P_N - P)\ell_f| + \epsilon .
\]

By Vapnik’s Lemma, it is sufficient to bound \( \sup_{f \in F} |(P_N - P)\ell_f| \) to prove the theorem. To obtain the result in expectation, we use the following lemma.

Lemma 9 (Pisier-Massart Lemma). Let \( U_1, \ldots, U_M \) denote random variables such that

\[
\forall s > 0, i \in \{1, \ldots, M\}, \quad \log \mathbb{E}[e^{sU_i}] \leq \frac{s^2 \sigma^2}{2} .
\]

Then

\[
\mathbb{E}[\max_{i=1,\ldots,M} U_i] \leq \sigma \sqrt{2 \log(M)} .
\]

Proof. By Jensen’s inequality, for any \( s > 0 \),

\[
\mathbb{E}[\max_{i=1,\ldots,M} U_i] = \frac{1}{s} \mathbb{E}[\log(\max_{i \in \{1,\ldots,M\}} e^{sU_i})] \leq \frac{1}{s} \log(\mathbb{E}[\max_{i \in \{1,\ldots,M\}} e^{sU_i}]) .
\]

Now, as the random variables \( e^{sU_i} \) are all nonnegative,

\[
\mathbb{E}[\max_{i \in \{1,\ldots,M\}} e^{sU_i}] \leq \mathbb{E}\left[ \sum_{i \in \{1,\ldots,M\}} e^{sU_i} \right] \leq M e^{s\sigma^2/2} .
\]
It follows that, for any $s > 0$,
\[
\mathbb{E}\left[ \max_{i=1,\ldots,M} U_i \right] \leq \frac{2\log M + s^2\sigma^2}{2s}
\]
In particular, for $s = \sqrt{2\log M/\sigma}$, this yields the result. □

From equation (1.4), the $2M$ random variables
\[
(P_N - P)\ell_f, \ (P - P_N)\ell_f, \ f \in F,
\]
have log-Laplace transform bounded from above by
\[
\frac{s^2}{2} \sum_{i=1}^{N} \frac{1}{4N^2} = \frac{s^2}{2} \frac{1}{4N}.
\]
It follows therefore from Pisier-Massart’s Lemma that
\[
\mathbb{E}[\sup_{f \in F}|(P_N - P)\ell_f|] \leq \frac{1}{2} \sqrt{\frac{\log 2M}{N}}.
\]
The oracle inequality in expectation in the theorem follows therefore from Vapnik’s lemma.

Applying Hoeffding’s theorem, we get that, for any $f \in F$ and any $t > 0$,
\[
\mathbb{P}((P_N - P)\ell_f > t) \leq e^{-2Nt^2}, \quad \mathbb{P}((P - P_N)\ell_f > t) \leq e^{-2Nt^2}.
\]
Applying a union bounds shows that,
\[
\forall t > 0, \quad \mathbb{P}\left(\sup_{f \in F}|(P_N - P)\ell_f| > t\right) \leq 2Me^{-2Nt^2}.
\]
Choosing $t$ solving the equation $2Me^{-2Nt^2} = \delta$ shows that
\[
\forall \delta \in (0, 1), \quad \mathbb{P}\left(\sup_{f \in F}|(P_N - P)\ell_f| > \sqrt{\frac{\log(2M/\delta)}{2N}}\right) \leq \delta.
\]
The proof is concluded thanks to Vapnik’s lemma. □

1.6 Benett’s and Bernstein’s inequalities

The rate of convergence in our first results $\sqrt{1/N}$ can sometimes be improved into “fast rates” $(1/N)^{\alpha}$ for some $\alpha \in [1/2, 1]$. The first ingredient to show this is to refine Hoeffding’s inequality to obtain a first order term scaling with the variance rather than the sup norm of the random variables in the empirical mean. This section presents Benett’s and Hoeffding’s inequalities which achieve this goal.
CHAPTER 1. EMPIRICAL RISK MINIMIZATION

**Theorem 10.** Let $Z_1, \ldots, Z_N$ denote independent random variables and let $f_1, \ldots, f_N$ denote measurable functions $f_i : \mathcal{Z} \to (-\infty, b]$. Let $\sigma^2 \geq (1/N) \sum_{i=1}^N \text{Var}(f_i(Z_i))$ and $h(u) = (1 + u) \log(1 + u) - u$. Then,

$$
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{N} \sum_{i=1}^N f_i(Z_i) - \mathbb{E}[f_i(Z_i)] > t \right) \leq e^{-N \frac{\sigma^2}{b^2} h\left(\frac{bt}{2}\right)}.
$$

**Proof.** One can assume, without loss of generality, that $b = 1$ and that $\mathbb{E}[f_i(Z_i)] = 0$. The proof is based on the remark that the function $\varphi$ is non-decreasing, where

$$
\forall u \in \mathbb{R}, \quad \varphi(u) = \frac{e^u - 1 - u}{u^2}.
$$

It follows that, for any $s > 0$,

$$
e^{sf_i(Z_i)} - 1 - sf_i(Z_i) \leq s^2 f_i^2(Z_i)(e^s - 1 - s).
$$

Taking the expectation on both sides yields

$$
\frac{1}{N} \sum_{i=1}^N \log(\mathbb{E}[e^{sf_i(Z_i)}]) \leq \log(1 + s^2 \sigma^2(e^s - 1 - s)) \leq s^2 \sigma^2(e^s - 1 - s).
$$

The result follows by applying Chernoff’s bound with $s = \log(1 + t/\sigma^2)$. \qed

**Remark 11.** Direct computations show that

$$h(u) \geq \frac{u^2}{2(1 + u)}.
$$

It follows that

$$
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{N} \sum_{i=1}^N f_i(Z_i) - \mathbb{E}[f_i(Z_i)] > t \right) \leq e^{-N \frac{\sigma^2}{b^2} h\left(\frac{bt}{2}\right)}.
$$

Equivalently,

$$
\forall t > 0, \quad \mathbb{P}\left( \frac{1}{N} \sum_{i=1}^N f_i(Z_i) - \mathbb{E}[f_i(Z_i)] > \sqrt{\frac{2\sigma^2 t}{N} + \frac{2bt}{N}} \right) \leq e^{-t}.
$$

This last inequality, known as Bernstein’s inequality, shows what we claimed at the beginning of the section: the first order term in the deviations of the empirical mean scales as a variance term $\sqrt{\sigma^2 t/N}$ rather than a sup-norm term $\sqrt{(b-a)^2 t/N}$ in Hoeffding’s inequality.
1.7 Fast rates

To prove faster rates of convergence, besides refined deviation inequalities, the following Bernstein condition will be needed. There exist $C > 0$ and $\alpha \in [0, 1]$ such that

$$\forall f \in F, \quad \text{Var}(\ell_f - \ell_{f_F}) \leq C (P[\ell_f - \ell_{f_F}])^\alpha .$$

(1.7)

The first part of this section shows that the Bernstein condition implies faster rates of convergence for the ERM than $1/N$. The second part presents some assumptions under which this condition is satisfied.

1.7.1 Fast rates under Bernstein’s condition

Theorem 12. Assume that (1.7) holds, then the $\epsilon$-ERM $\hat{f}_{erm}$ over the finite dictionary $F = \{f_1, \ldots, f_M\}$ satisfies

$$\mathbb{P}(\mathcal{E}_{\ell,F}(\hat{f}_{erm}) \leq \gamma_{\alpha,C} \left( \frac{\log(M/\delta)}{N} \right)^{\frac{1}{2-\alpha}} + 4 \frac{\log(M/\delta)}{N} + \epsilon ) \geq 1 - \delta .$$

Here, one can choose $\gamma_{\alpha,C} = (2 - \alpha)\alpha^{\frac{\alpha}{2}}(2C)^{\frac{1}{2-\alpha}}$.

Proof. We prove the theorem for the actual $\hat{f}_{erm}$, with $\epsilon = 0$. The extension to $\epsilon > 0$ is left to the reader. Recall that, from the proof of Vapnik’s Lemma,

$$\mathcal{E}_{\ell,F}(\hat{f}_{erm}) \leq (PN - P)[\ell_{f_F} - \ell_{\hat{f}_{erm}}] .$$

(1.8)

By Bernstein’s inequality, for any $f \in F$, with probability at least $1 - \delta/M$,

$$(PN - P)[\ell_{f_F} - \ell_f] \leq \sqrt{\frac{2\text{Var}(\ell_{f_F} - \ell_f) \log(M/\delta)}{N}} + \frac{2\log(M/\delta)}{N} .$$

Applying a union bound and Bernstein’s condition (1.7), this shows that, with probability at least $1 - \delta$, for any $f \in F$,

$$(PN - P)[\ell_{f_F} - \ell_f] \leq \sqrt{\frac{2C\mathcal{E}_{\ell,F}^\alpha(f) \log(M/\delta)}{N}} + \frac{2\log(M/\delta)}{N} .$$

By Minkowsky’s inequality, on the same event,

$$(PN - P)[\ell_{f_F} - \ell_f] \leq \frac{1}{2} \mathcal{E}_{\ell,F}(f) + \frac{(2 - \alpha)\alpha^{\frac{\alpha}{2}}C^{\frac{1}{2-\alpha}}}{2^{\frac{1}{2-\alpha}}} \left( \frac{\log(M/\delta)}{N} \right)^{1/(2-\alpha)}$$

$$+ \frac{2\log(M/\delta)}{N} .$$

$\square$
1.7.2 Examples satisfying the Bernstein’s assumption

In this section, we assume that $f^* \in F$, so $f^*_F = f^*$ and $\mathcal{E}_{\ell,F}(f) = \mathcal{E}(f)$. Remark that $|\ell_f(X,Y) - \ell_{f^*}(X,Y)| \leq 1_{f(X) \neq f^*(X)}$, therefore

$$\text{Var}(\ell_f - \ell_{f^*}) = \text{Var}(\ell_f - \ell_{f^*}) \leq P(f(X) \neq f^*(X)) = \mathcal{E}(f). \quad (1.9)$$

Recall the representation of the excess risk provided in Theorem 2.

It transpires from this representation that a natural way to bound $\text{Var}(\ell_f - \ell_{f^*})$ by $\mathcal{E}(f)$ is to assume that $\eta(X) = 1/2$ means that $X$ does not bring information on $Y$ or that the classification task. Therefore, assuming that this is not the case corresponds intuitively to assuming that the classification task should be easier. The purpose of this section is to show that, indeed, “margin” assumptions help the ERM to improve its rates of convergence.

The noiseless case Assume first that we are in the ideal case where $\eta(X) = Y$, that is all the relevant information on $Y$ is contained in $X$. In this case, $\eta(X) \in \{0, 1\}$ so, by the representation theorem

$$\mathcal{E}(f) = \mathbb{E}[|2\eta(X) - 1|1_{f(X) \neq f^*(X)}] = \mathbb{P}(f(X) \neq f^*(X)).$$

By (1.9), the “ideal” Bernstein’s condition holds with $C = 1$, $\alpha = 1$. From Theorem 12

$$\forall \delta \in (0, 1), \quad \mathbb{P}\left(\mathcal{E}(f) \leq \frac{6 \log(M/\delta)}{N}\right) \geq 1 - \delta.$$  

Massart’s margin condition A closely related assumption was proposed by Massart. Instead of assuming that all the information is contained in $X$, which may be restrictive, Massart suggested that it is sufficient to bound away $\eta$ from 1/2 to check Bernstein’s condition. The following condition is also known as the “hard” margin assumption: there exists $\gamma \in (0, 1/2]$ such that, almost surely

$$|\eta(X) - 1/2| \geq \gamma.$$  

Remark that the noiseless case corresponds to $\gamma = 1/2$. Under Massart’s condition, it holds that

$$\mathcal{E}(f) = \mathbb{E}[|2\eta(X) - 1|1_{f(X) \neq f^*(X)}] \geq 2\gamma \mathbb{P}(f(X) \neq f^*(X)).$$
1.7. FAST RATES

By (\ref{eq:bernstein}), Bernstein’s condition holds with $C = 1/(2\gamma)$, $\alpha = 1$. From Theorem \ref{thm:bernstein},

$$\forall \delta \in (0, 1), \quad \mathbb{P}\left(\mathcal{E}(f) \leq \left(\frac{1}{\gamma} + 4\right)\frac{\log(M/\delta)}{N}\right) \geq 1 - \delta.$$

**Mammen-Tsybakov’s margin condition**  Mammen and Tsybakov further relaxed the hard margin condition and show that it is actually sufficient to bound the probability that $\eta(X)$ lies in the neighborhood of $1/2$ of size $t$ to check Bernstein’s condition. They introduced the following “soft” margin assumption: there exist $t_0$, $C_0$ and $\nu \in (0, 1)$ such that, for any $t \leq t_0$,

$$\mathbb{P}(|\eta(X) - 1/2| \leq t) \leq C_0 t^{\frac{\nu}{1-\nu}}.$$

Fix some $t \leq t_0$, under Mammen-Tsybakov’s condition

$$\mathcal{E}(f) = \mathbb{E}\left[|2\eta(X) - 1|1_{f(X) \neq f^*(X)}\right] \geq \mathbb{E}\left[|2\eta(X) - 1|1_{(\eta(X) - 1/2 > t)}1_{(f(X) \neq f^*(X))}\right] \geq 2t \mathbb{P}(\eta(X) - 1/2 > t \cap f(X) \neq f^*(X)) \geq 2t \mathbb{P}(f(X) \neq f^*(X)) - \mathbb{P}(|\eta(X) - 1/2| \leq t) \geq 2t \mathbb{P}(f(X) \neq f^*(X)) - 2C_0 t^{\frac{1}{1-\nu}}.$$

Choose now $t = t_0 \mathcal{E}(f)^{1-\nu}/(1 \lor 4C_0)$ so $t \leq t_0$ and $2C_0 t^{\frac{1}{1-\nu}} \leq \mathcal{E}(f)/2$. Therefore

$$\frac{2t_0}{1 \lor 4C_0} \mathcal{E}(f)^{1-\nu} \mathbb{P}(f(X) \neq f^*(X)) \leq \frac{3\mathcal{E}(f)}{2},$$

that is, by (\ref{eq:bernstein}),

$$\text{Var}(\ell_f - \ell_{f^*}) \leq \frac{3(1 \lor 4C_0)}{4t_0} \mathcal{E}(f)^\nu.$$

Bernstein’s condition holds with $C = 3(1 \lor 4C_0)/(4t_0)$, $\alpha = \nu$. From Theorem \ref{thm:bernstein},

$$\forall \delta \in (0, 1), \quad \mathbb{P}\left(\mathcal{E}(f) \leq \gamma_{C_0, t_0} \left(\frac{\log(M/\delta)}{N}\right)^{\frac{1}{1-\nu}}\right) \geq 1 - \delta.$$
Chapter 2
Learning with infinite resources

2.1 The general framework

Assume that $Z_1, \ldots, Z_N$ are couples $Z_i = (X_i, Y_i)$ taking values in $\mathcal{X} \times [-1, 1]$. Assume that $F$ is a set of functions $f : \mathcal{X} \rightarrow [-1, 1]$. Assume that $\ell_f(z) \in [0, 1]$. The risk of $f \in F$ is $R(f) = P\ell_f$. Oracles $\tilde{f}$ are defined by

$$\tilde{f} \in \arg\min_{f \in F} R(f).$$

Empirical risk minimizers (ERM) $\hat{f}_{\text{erm}}$ are defined by

$$\hat{f}_{\text{erm}} \in \arg\min_{f \in F} P_N \ell_f.$$

From Vapnik’s Lemma, risk bounds for the ERM follow from upper bounds on $\sup_{f \in F} |(P_N - P) \ell_f|.$

2.1.1 Example: Lipschitz loss and linear functionals

All along this section, besides binary classification, general results will be illustrated on the following example. The importance of this example will become clear in the following chapter.

Assume that the loss $\ell$ satisfies $\ell_f(x, y) = c(f(x), y)$, for all $f : \mathcal{X} \rightarrow \mathbb{R}$, $x \in \mathcal{X}$ and $y \in \mathcal{Y}$ and that the cost function $c$ is (uniformly over $\mathcal{Y}$) Lipschitz in its first argument

$$\exists L > 0 : \forall u, v \in \mathbb{R}, \forall y \in \mathcal{Y}, \quad c(u, y) - c(v, y) \leq L|u - v|. \quad (2.1)$$

Let $B_\infty = \{a \in \mathbb{R}^d : \|a\|_\infty = \max_{i \in \{1, \ldots, d\}} |a_i| \leq 1\}$. Assume furthermore that $\mathcal{X} = B_1 = \{x \in \mathbb{R}^d : \|x\|_1 = \sum_{i=1}^d |x_i| \leq 1\}$ and that

$$F = \{f = f_a, \ a \in B_\infty\}, \quad \text{where} \quad \forall a, x \in \mathbb{R}^d, \quad f_a(x) = a^T x.$$
2.2 Azuma-Hoeffding’s inequality

Let $\mathcal{F}_t$ denote a filtration on $\mathcal{X}$ and $\Delta_t$ denote an adapted martingale difference, that is $\Delta_t$ is $\mathcal{F}_t$-measurable and satisfies

$$\mathbb{E}[\Delta_t|\mathcal{F}_s] = 0, \quad \forall s < t.$$ 

**Theorem 13.** Assume that, for any $t$, there exist $A_t, B_t, \mathcal{F}_{t-1}$ measurable such that, almost surely $A_t \leq \Delta_t \leq B_t$. Then

$$\forall u > 0, \quad \mathbb{P}\left(\frac{1}{N}\sum_{t=1}^{N} \Delta_t > u\right) \leq e^{-2N^2u^2/\sum_{t=1}^{N} \|B_t - A_t\|_\infty^2}.$$ 

**Proof.** By the Chernoff bound, it is sufficient to bound the Fenchel Legendre transform $\psi^*_{\frac{1}{N}\sum_{t=1}^{N} \Delta_t}(s) = \log \mathbb{E}[e^{(s/N)\sum_{t=1}^{N} \Delta_t}] = \log \mathbb{E}[e^{(s/N)\sum_{t=1}^{N-1} \Delta_t}\mathbb{E}[e^{(s/N)\Delta_N}|\mathcal{F}_{N-1}]]$.

Conditionally on $\mathcal{F}_{N-1}$, $\Delta_N$ is a random variable taking values in $[A_N, B_N]$, therefore, by Hoeffding’s lemma,

$$\mathbb{E}[e^{(s/N)\Delta_N}|\mathcal{F}_{N-1}] \leq e^{s^2(B_N - A_N)^2/(8N^2)} \leq e^{s^2\|B_N - A_N\|_\infty^2/(8N^2)}.$$ 

It follows that

$$\psi^*_{\frac{1}{N}\sum_{t=1}^{N} \Delta_t}(s) \leq \log \mathbb{E}[e^{(s/N)\sum_{t=1}^{N-1} \Delta_t}] + \frac{\|B_N - A_N\|_\infty^2 s^2}{8N^2}.$$ 

By recurrence, we get

$$\psi^*_{\frac{1}{N}\sum_{t=1}^{N} \Delta_t}(s) \leq \frac{s^2}{2} \frac{\sum_{t=1}^{N} \|B_t - A_t\|_\infty^2 s^2}{4N^2}.$$ 

Together with (1.5), this inequality shows that the Fenchel-Legendre can be bounded from below by

$$\psi^*_{\frac{1}{N}\sum_{t=1}^{N} \Delta_t}(u) \geq \frac{2N^2u^2}{\sum_{t=1}^{N} \|B_t - A_t\|_\infty^2}.$$ 

The proof is concluded by the Chernoff bound. \qed
2.3. THE BOUNDED DIFFERENCE INEQUALITY

2.3 The bounded difference inequality

In this section \( n \) is an integer, \( c \in \mathbb{R}_+^n \), \( g : \mathbb{Z}^n \to \mathbb{R} \) is a measurable function and \( Z = (Z_1, \ldots, Z_n) \) is a vector of independent random variables taking values in \( \mathbb{Z} \). Start with a definition.

Definition 14. The function \( g \in \mathbb{BD}(c) \) if

\[
\forall x, y \in \mathbb{R}^n, \quad |g(x) - g(y)| \leq \sum_{i=1}^n c_i \mathbf{1}_{\{x_i \neq y_i\}}.
\]

Theorem 15 (Bounded Difference Inequality). Assume that \( g \in \mathbb{BD}(c) \), then

\[
\forall t > 0, \quad \mathbb{P}(g(Z) > \mathbb{E}[g(Z)] + t) \leq e^{-2t^2/\sum_{i=1}^n c_i^2}.
\]

Proof. The proof relies on Azuma-Hoeffding’s inequality. Let \( \mathcal{F}_t \) denote the sigma-algebra induced by \( X_1, \ldots, X_t, \mathcal{F}_0 = \{\emptyset, \Omega\} \) and let

\[
\Delta_t = \mathbb{E}[g(Z)|\mathcal{F}_t] - \mathbb{E}[g(Z)|\mathcal{F}_{t-1}].
\]

\( \Delta_t \) are by construction martingale increments with respect to the filtration \( \mathcal{F}_t \). Let

\[
B_t = \mathbb{E}[\sup_{z \in \mathbb{Z}} g(Z_1, \ldots, Z_{t-1}, z_t, Z_{t+1}, \ldots, Z_N) - g(Z)|\mathcal{F}_{t-1}]
\]

\[
A_t = \mathbb{E}[\inf_{z \in \mathbb{Z}} g(Z_1, \ldots, Z_{t-1}, z_t, Z_{t+1}, \ldots, Z_N) - g(Z)|\mathcal{F}_{t-1}].
\]

By construction, \( \Delta_t \in [A_t, B_t] \) almost surely and, since \( g \in \mathbb{BD}(c) \),

\[
\|B_t - A_t\|_\infty \leq c_i.
\]

The theorem follows therefore from a direct application of Azuma-Hoeffding’s inequality.

2.4 Symmetrization principle

From Vapnik’s lemma, the risk of ERM \( f_{\text{erm}} \) will be bounded by \( \sup_{f \in F} |(P_N - P)\ell_f| \). By the bounded difference inequality, it will be enough to bound

\[
\mathbb{E}[\sup_{f \in F} (P_N - P)\ell_f].
\]

Bounding this quantity is hard in general. A common step to many methods though is to apply the following symmetrization trick.
Lemma 16 (Symmetrization Lemma). For any loss functions $\ell$, any set of parameters $F$, 

$$
\mathbb{E} \left[ \sup_{f \in F} |(P_N - P)\ell_f| \right] \leq 2 \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(Z_i) \right| \right] 

\leq 2 \sup_{z \in \mathcal{Z}} \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(z_i) \right| \right],
$$

where $\epsilon_1, \ldots, \epsilon_N$ are i.i.d. Rademacher random variables.

Proof. Let $\mathcal{D}'_N = (Z'_1, \ldots, Z'_N)$ denote an independent copy of $\mathcal{D}_N$, let $P'_N$ denote the empirical process based on $\mathcal{D}'_N$, that is $P'_N g = N^{-1} \sum_{i=1}^{N} g(Z'_i)$ for any function $g : \mathcal{Z} \rightarrow \mathbb{R}$. A key remark underlying the symmetrization principle is that

$$\forall f \in F, \quad P\ell_f = \mathbb{E}[P'_N \ell_f | \mathcal{D}_N].$$

By Jensen’s inequality, it follows that

$$\mathbb{E} \left[ \sup_{f \in F} |(P_N - P)\ell_f| \right] = \mathbb{E} \left[ \sup_{f \in F} |P_N \ell_f - \mathbb{E}[P'_N \ell_f | \mathcal{D}_N]| \right]$$

$$\leq \mathbb{E} \left[ \sup_{f \in F} |P_N \ell_f - P'_N \ell_f| | \mathcal{D}_N | \right]$$

$$\leq \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} (\ell_f(Z_i) - \ell_f(Z'_i)) \right| \right].$$

As the vectors $(\ell_f(Z_i) - \ell_f(Z'_i))_{i \in \{1, \ldots, N\}, f \in F}$ and $(\epsilon_i (\ell_f(Z_i) - \ell_f(Z'_i)))_{i \in \{1, \ldots, N\}, f \in F}$ have the same distribution, it follows that

$$\mathbb{E} \left[ \sup_{f \in F} |(P_N - P)\ell_f| \right] \leq \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i (\ell_f(Z_i) - \ell_f(Z'_i)) \right| \right]$$

$$\leq 2 \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(Z_i) \right| \right].$$ 

\square

2.5 Application to learning theory

Recall that Vapnik’s Lemma implies that the risk of the Empirical Risk Minimizer $\hat{f}_{\text{erm}}$ is bounded from above by

$$\mathcal{E}_{\ell, F}(\hat{f}_{\text{erm}}) \leq 2 \sup_{f \in F} |(P_N - P) f|.$$
As \( \ell_f \in [0,1] \) the functional sup\(_{f \in F} |(P_N - P)\ell_f| \) belongs to \( \mathbb{BD}(\mathbf{e}) \), with \( c_i = 1/N \) for any \( i \in \{1, \ldots, N\} \). Therefore, from the bounded difference inequality, for any \( \delta \in (0,1) \), with probability larger than \( 1 - \delta \),

\[
\sup_{f \in F} |(P_N - P)\ell_f| \leq \mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] + \sqrt{\frac{\log(1/\delta)}{2N}}.
\]

By the symmetrization lemma, it follows that

\[
\sup_{f \in F} |(P_N - P)\ell_f| \leq 2\mathcal{R}_N(\ell, F) + \sqrt{\frac{\log(1/\delta)}{2N}},
\]

where \( \mathcal{R}_N(\ell, F) \) denotes the Rademacher complexity of \( F \) with respect to the loss \( \ell \), defined by

\[
\mathcal{R}_N(\ell, F) = \sup_{z \in \mathbb{Z}^N} \mathbb{E}\left[ \sup_{f \in F} \frac{1}{N} \sum_{i=1}^N \epsilon_i \ell_f(z_i) \right].
\]

### 2.6 The finite case

**Lemma 17.** If \( F \) is finite, then

\[
\mathcal{R}_N(\ell, F) \leq \sqrt{\frac{2\log(2|F|)}{N}}.
\]

**Proof.** Consider the set \( B \) of the \( |F| \) vectors \( (\ell_f(z_i))_{1 \leq i \leq N} \in \mathbb{R}^N \). As their coordinates belong to \([0,1] \), \( \Delta(B) \leq \sqrt{N} \). Therefore, the proof terminates with Lemma 18. \( \square \)

**Lemma 18.** Let \( B \) denote a finite subset of vectors in \( \mathbb{R}^N \), and let \( \Delta(B) = \max_{b \in B} \|b\| \). Then

\[
N\mathcal{R}_N(B) = \mathbb{E}\left[ \max_{b \in B} \left| \sum_{i=1}^N \epsilon_i b_i \right| \right] \leq \Delta(B) \sqrt{2\log(2|B|)}.
\]

**Proof.** For any \( b \in B \), let \( Z_b = \sum_{i=1}^N \epsilon_i b_i \). As \(-|b_i| \leq \epsilon_i b_i \leq |b_i|\), Hoeffding’s Lemma ensures that

\[
\forall s > 0, \quad \log(\mathbb{E}[e^{sZ_b}]) = \sum_{i=1}^N \log(\mathbb{E}[e^{s\epsilon_i b_i}]) \leq \frac{s^2\|b\|^2}{2} \leq \frac{s^2\Delta(B)^2}{2}.
\]

Of course the bound holds for the random variable \(-Z_b\). By Pisier-Massart’s Lemma, it follows that

\[
\mathbb{E}\left[ \max_{b \in B} |Z_b| \right] \leq \Delta(B) \sqrt{2\log(2|B|)}.
\]

\( \square \)
2.7 Vapnik-Chervonenkis theory

The goal of this section is to bound $\mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|]$ in the binary classification setting. For any $f \in F$, let $A_f$ denote the measurable set such that $\{Z \in A_f\} = \{Y \neq f(X)\}$. Let $A_F = \{A_f \in f \in F\}$. Introduce the following Rademacher complexity.

**Definition 19.** The Rademacher complexity of the collection of sets $\mathcal{A}$ in the space $\mathcal{Z}$ is defined as

$$R_N(\mathcal{A}) = \sup_{z_1, \ldots, z_N \in \mathcal{Z}} \mathbb{E} \left[ \sup_{A \in \mathcal{A}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i 1_{\{z_i \in A\}} \right| \right].$$

It follows from the symmetrization Lemma that

$$\mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] \leq 2R_N(A_F).$$

Introduce, for any $N \geq 1$, the set of vectors

$$T_N(z) = \{(1_{\{z_1 \in A_f\}}, \ldots, 1_{\{z_N \in A_f\}})^T, f \in F\}, \quad z = (z_1, \ldots, z_N) \in \mathcal{Z}^N.$$

For any $z \in \mathcal{Z}^N$, $|T_N(z)| \leq 2^N$.

**Definition 20.** $F$ shatters $z \in \mathcal{Z}^N$ if $|T_N(z)| = 2^N$. The shattering coefficients of $F$ is the sequence $(S_n(F))_{n \geq 1}$, where, for any $n \geq 1$,

$$S_n(F) = \sup_{z \in \mathcal{Z}^n} \{|T_n(z)|\}.$$

The VC dimension of $F$ is the largest integer $d$ such that $S_d(F) = 2^d$. We write $\text{VC}(F) = d$.

In words, there exists a set of $d$ points $z_1, \ldots, z_d$ in $\mathcal{Z}$ such that $F$ shatters $z = (z_1, \ldots, z_d)$ but there is not a single set of $d + 1$ points that is shattered by $F$. In particular therefore, $F$ does not shatter a single set of $d' > d$ points and the VC dimension is well defined.

**Exercise** Check that the VC dimension of half lines is 2.

### 2.7.1 VC inequality

**Theorem 21.** Assume that $F$ has VC dimension $d$, then

$$\mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] \leq 2\sqrt{\frac{2d \log(2eN/d)}{N}}.$$
Proof. Recall that, by the symmetrization trick, it holds
\[ \mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] \leq 2R_N(A_F) = \sup_{z \in \mathbb{Z}^N} \mathcal{R}_N(T_N(z)) . \]

For any \( z \in \mathbb{Z}^N \), any vector in \( T_N(z) \) has coordinates in \( \{0, 1\} \), in particular therefore \( \Delta(T_N(z)) \leq \sqrt{N} \). Moreover, by definition, for any \( z \in \mathbb{Z}^N \), \(|T_N(z)| \leq S_N(F)\), therefore, by Lemma 13,
\[ \mathbb{E}[\sup_{f \in F} |(P_N - P)\ell_f|] \leq 2 \sup_{z \in \mathbb{Z}^N} \mathcal{R}_N(T_N(z)) \leq \sqrt{\frac{2 \log(2S_N(F))}{N}} . \]

The following lemma concludes the proof of the theorem.

**Lemma 22** (Sauer’s Lemma). If \( \text{VC}(F) = d \), then for all \( N \geq 1 \),
\[ S_N(F) \leq \sum_{k=0}^d \binom{N}{k} . \]

For any \( N \geq d \), in particular,
\[ S_N(F) \leq \left( \frac{eN}{d} \right)^d . \]

Together with the BDI, Theorem 21 implies the following corollary.

**Theorem 23.** Let \( F \) denote a set of classifiers with VC dimension \( d \) and let \( \hat{f}_{\text{erm}} \) denote the empirical risk minimizer over \( F \). Then, for any \( \delta \in (0, 1) \),
\[ \mathcal{E}_{\ell,F}(\hat{f}_{\text{erm}}) \leq 4 \sqrt{\frac{2d \log(2eN/d)}{N}} + \sqrt{\frac{2 \log(1/\delta)}{N}} . \]

**Remark 24.** The set \( F \) may have finite VC dimension and be infinite (think about the half lines). Therefore, Theorem 23 extends the slow rates to possibly infinite dictionaries.

### 2.8 Covering numbers

**Definition 25.** Let \( K \) denote a set endowed with a metric \( d \) and let \( \eta > 0 \). An \( \eta \)-net of \((K,d)\) is set \( V \) such that, for any \( f \in F \), there exists \( g \in V \) such that \( d(f,g) \leq \eta \). The covering numbers of \( K,d \) are defined by
\[ N(K,d,\eta) = \inf\{|V| : V \text{ is an } \eta \text{-net of } (K,d)\} . \]
Define, for any \( f \) and \( g \) in \( F \), and any \( p \geq 1 \),
\[
d^p_p(f, g) = \left( \frac{1}{N} \sum_{i=1}^{N} |\ell_f(z_i) - \ell_g(z_i)|^p \right)^{1/p}.
\]

**Theorem 26.** For any \( z \in \mathcal{Z}^N \),
\[
\mathcal{R}_N^z(F) = \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(z_i) \right| \right] \leq \inf_{\eta > 0} \left\{ \eta + \sqrt{\frac{2 \log(2N(F, d_1^Z, \eta))}{N}} \right\}.
\]

**Proof.** Fix \( z \in \mathcal{Z}^N \) and \( \eta > 0 \). Let \( V \) denote an \( \eta \)-net of \((F, d_1^Z)\) such that \(|V| = N(F, d_1^Z, \eta)\). For any \( f \in F \), denote by \( \pi_V(f) \in V \) an element such that \( d(f, \pi_V(f)) \leq \eta \). By the triangular inequality,
\[
\mathcal{R}_N^z(F) \leq \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i (\ell_f(z_i) - \ell_{\pi_V(f)}(z_i)) \right| \right] + \mathbb{E} \left[ \sup_{f \in F} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_{\pi_V(f)}(z_i) \right| \right] \\
\leq \epsilon + \mathbb{E} \left[ \max_{f \in V} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i \ell_f(z_i) \right| \right].
\]

The proof terminates thanks to Lemma 17. \( \square \)

In the main example of Section 2.1.1, Theorem 26 yields the following result.

**Theorem 27.** Consider the framework of Section 2.1.1. Then for any \( \delta \in (0, 1) \),
\[
\mathbb{P} \left( \mathcal{E}_{L,F}(\hat{f}_{erm}) \leq 8 \sqrt{\frac{d \log(2eLN/d)}{N}} + \sqrt{\frac{2 \log(1/\delta)}{N}} \right) \geq 1 - \delta.
\]

**Proof.** Start with a basic lemma.

**Lemma 28.** For any \( f \) in \( F \), denote by \( a_f \in B_\infty \) the vector such that, for all \( x \in \mathcal{X} \), \( f(x) = a_f^T x \). For any \( f, g \in F \) and any \( z \in \mathcal{Z}^N \), and any \( p \geq 1 \)
\[
d^p_p(f, g) = \left( \frac{1}{N} \sum_{i=1}^{N} |\ell_f(z_i) - \ell_g(z_i)|^p \right)^{1/p} \\
\leq L \left( \frac{1}{N} \sum_{i=1}^{N} |(a_f - a_g)^T x_i|^p \right)^{1/p} \leq L \|a_f - a_g\|_\infty.
\]
2.8. COVERING NUMBERS

The proof is straightforward. To bound the covering number of \((F, d_1^z)\), it is therefore sufficient to bound the covering number of \((B_\infty, d_\infty)\), where 
\[ d_\infty(a, b) = \|a - b\|_\infty, \] 
for all \(a, b \in \mathbb{R}^d\). This is done in the following lemma.

**Lemma 29.** For any \(\eta > 0\), there exists an \(\eta\)-net of \((B_\infty, d_\infty)\) with cardinality \((2/\eta)^d\).

To build an \(\eta\)-net of \((F, d_1^z)\), it is therefore enough to build an \((\eta/L)^d\)-net of \((B_1, d_1)\). Let \(V_\eta\) denote the set of all vectors in \(\mathbb{R}^d\) with coordinates taking values in the grid 
\[ G_\eta = \{\eta/2 + k\eta, k \in \{-k_0, \ldots, k_0 - 1\}\} \subset [-1, 1]. \]

\(V_\eta\) has cardinality \((2k_0)^d \leq (2/\eta)^d\) and, for any \(a \in B_\infty\), there exist \(v \in V_\eta\) such that 
\[ \|a - v\|_\infty = \max_{i \in \{1, \ldots, d\}} \min_{g \in G} |a_i - g| \leq \eta. \]

Hence, \(V_\eta\) is an \(\eta\)-net of \((B_\infty, d_\infty)\). From \((2.3)\), it follows that the set \(\{f_v, v \in V_{\eta/L}\}\) is an \((\eta/L)^d\)-net of \((F, d_1^z)\). It follows that \(N(F, d_1^z, \eta) \leq (2L/\eta)^d\). As this holds for any \(z \in \mathcal{G}_N\), it follows from Theorem 26 that the Rademacher complexity of \(F\) can be bounded from above by 
\[ \mathcal{R}_N(\ell, F) \leq \inf_{\eta > 0} \left\{ \eta + \sqrt{\frac{2d\log(4L/\eta)}{N}} \right\} \]
\[ \leq \sqrt{\frac{2d}{N}} \left( 1 + \sqrt{\log(2LN/d)} \right) \]
\[ \leq 2 \sqrt{\frac{d\log(2eLN/d)}{N}}. \]

The first inequality follows by taking \(\eta = \sqrt{2d/N}\) and the second comes from 
\[ 1 + \sqrt{x} \leq \sqrt{2(1 + x)}. \]

Using \((2.2)\) and the symmetrization lemma, it follows that, for all \(\delta \in (0, 1), \)
\[ \mathbb{P}\left( \sup_{f \in F} |(P_N - P)\ell_f| \leq 4 \sqrt{\frac{d\log(2eLN/d)}{N}} + \sqrt{\frac{\log(1/\delta)}{2N}} \right) \geq 1 - \delta. \]

Finally, from Vapnik’s lemma, for all \(\delta \in (0, 1), \)
\[ \mathbb{P}\left( R(\hat{f}_{\text{erm}}) \leq R(\hat{f}) + 8 \sqrt{\frac{d\log(2eLN/d)}{N}} + \frac{2\log(1/\delta)}{N} \right) \geq 1 - \delta. \]

\[ \square \]

Theorem 27 establishes a risk bound for ERM in a case where the dictionary is infinite. This bound can be refined using a slightly sharper bound of the Rademacher complexity. The chaining method allowing this sharper analysis is at the core of the following section.
2.9 The chaining method

To establish the main result of this section, let

$$d^2_z(f, g) = \left( \frac{1}{N} \sum_{i=1}^{N} (\ell_f(z_i) - \ell_g(z_i))^2 \right)^{1/2}.$$  

Theorem 30. For any $z \in \mathbb{Z}^N$,

$$\mathcal{R}_N^*(F) \leq \inf_{\varepsilon > 0} \left\{ 4\varepsilon + \frac{12}{\sqrt{N}} \int_{\varepsilon}^{1} \sqrt{\log(N(F, d^2_z, t))} dt \right\}.$$  

Proof. Fix $z \in \mathbb{Z}^N$. For any $j \geq 1$, denote by $V_j$ a $2^{-j}$-net of $(F, d^2_z)$ with cardinality $N(F, d^2_z, 2^{-j})$. Let $f = (\ell_f(z_1), \ldots, \ell_f(z_N))^T$ and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_N)^T$ so

$$\mathcal{R}_N^*(F) = \frac{1}{N} \mathbb{E}[\sup_{f \in F} \varepsilon^T f]$$

For any $f \in F$, let $\pi_j(f) \in V_j$ such that $d^2_z(f, \pi_j(f)) \leq 2^{-j}$ and $f_j = (\ell_{\pi_j(f)}(z_1), \ldots, \ell_{\pi_j(f)}(z_N))^T$. Define $\pi_0(f)$, so, for any $f \in F$ and any $j_0$,

$$\varepsilon^T f = \varepsilon^T (f - f_{j_0}) + \sum_{j=1}^{j_0} \varepsilon^T (f_j - f_{j-1}).$$

In particular therefore,

$$\mathcal{R}_N^*(F) \leq \frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f - f_{j_0})|] + \sum_{j=1}^{j_0} \frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|].$$

First, by Cauchy-Schwarz inequality,

$$\frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f - f_{j_0})|] \leq \frac{\|\varepsilon\|_2 d^2_z(f, \pi_{j_0}(f))}{\sqrt{N}} \leq 2^{-j_0}.$$  

Second, to bound

$$\mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|],$$

notice first that there are at most $|V_j||V_{j-1}| = N(F, d^2_z, 2^{-j+1})N(F, d^2_z, 2^{-j})$ possibles values for the couples $(f_{j-1}, f_j)$. Next, by Lemma [??].

$$\frac{1}{N} \mathbb{E}[\sup_{f \in F} |\varepsilon^T (f_j - f_{j-1})|] \leq \max_{f \in F} \|f_j - f_{j-1}\|_2 \sqrt{2 \log(|V_j||V_{j-1}|)}.$$
Then
\[
\|f_j - f_{j-1}\|_2 = \sqrt{N}d^2_2(\pi_j(f) - \pi_{j-1}(f)) \leq 3\sqrt{N}2^{-j}.
\]
Substituting in the previous bound yields
\[
\frac{1}{N}E[\sup_{f \in F}|\varepsilon^T(f_j - f_{j-1})|] \leq 3(2^{-j})\sqrt{\frac{2\log(|V_j||V_{j-1}|)}{N}}.
\]
Summing up over \(j\) yields
\[
\sum_{j=1}^{j_0} \frac{1}{N}E[\sup_{f \in F}|\varepsilon^T(f_j - f_{j-1})|] \leq \sum_{j=1}^{j_0} 3(2^{-j})\sqrt{\frac{2\log(|V_j|)}{N}}
\]
\[
+ \sum_{j=1}^{j_0} 3(2^{-j})\sqrt{\frac{2\log(|V_{j-1}|)}{N}}
\]
\[
= \sum_{j=1}^{j_0} 6(2^{-j} - 2^{-j-1})\sqrt{\frac{2\log(|V_j|)}{N}}
\]
\[
+ \sum_{j=1}^{j_0} 3(2^{-j+1} - 2^{-j})\sqrt{\frac{2\log(|V_{j-1}|)}{N}}
\]
\[
\leq 9 \sum_{j=1}^{j_0} (2^{-j} - 2^{-j-1})\sqrt{\frac{2\log(|V_j|)}{N}}
\]
\[
= 9 \sum_{j=1}^{j_0} (2^{-j} - 2^{-j-1})\sqrt{\frac{2\log(N(F,d_2^2,2^{-j}))}{N}}.
\]
The function \(x \mapsto \sqrt{\log(N(F,d_2^2,x))}\) being non-increasing, this sum can be bounded from above by the integral
\[
\sum_{j=1}^{j_0} \frac{1}{N}E[\sup_{f \in F}|\varepsilon^T(f_j - f_{j-1})|] \leq \int_{2^{-j_0-1}}^{1/2} \sqrt{\log(N(F,d_2^2,x))}dx.
\]
Choosing \(j_0\) such that \(2^{-j_0-2} \leq \eta \leq 2^{-j_0-1}\) yields
\[
\mathcal{R}_N^*(F) \leq 4\eta + 9 \int_{\eta}^{1} \sqrt{\log(N(F,d_2^2,x))}dx.
\]

The advantage of this refined analysis will be clearer in the example of Section 2.1.1.
**Theorem 31.** Consider the framework of Section 2.1.1. There exists an absolute constant $C$ such that, for any $\delta \in (0, 1)$,

$$
P\left(\mathcal{E}(\hat{f}_{erm}) \leq C\left(\sqrt{\frac{d\log(eL)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right)\right) \geq 1 - \delta.
$$

**Proof.** Let us first compute $N(F, d^2_z, x)$ for any $x > 0$. Recall that $F = \{f_a, \ a \in B_\infty\}$. Fix $z \in \mathcal{Z}^N$, then

$$
d^2_z(f, g) = \left(\frac{L^2}{N} \sum_{i=1}^N [(a_f - a_g)^T x_i]^2\right)^{1/2} \leq L\|a_f - a_g\|_\infty.
$$

It follows therefore from Lemma 29 that

$$
N(F, d^2_z, x) \leq (2L/x)^d.
$$

Therefore, for any $\eta > 0$,

$$
\int_{\eta}^{1} \sqrt{\log(N(F, d^2_z, x))} dx \leq \int_{0}^{1} \sqrt{d\log(2L/x)} dx \leq \sqrt{d\log(2L)} + C\sqrt{d}.
$$

The Chaining theorem and (2.2) imply therefore that

$$
\sup_{f \in F} |(P_N - P)\ell_f| \leq C\left(\sqrt{\frac{d\log(eL)}{N}} + \sqrt{\frac{\log(1/\delta)}{N}}\right).
$$

Vapnik’s Lemma concludes the proof. \qed
Chapter 3
Convex relaxation

Solving the optimization problem defining the ERM
\[ \hat{f}_{\text{erm}} = \arg\min_{f \in F} P_N 1_{\{Y \neq f(X)\}} \]
is computationally demanding. In fact, for most classes \( F \), it simply cannot be done. The basic idea of this chapter is to replace the non-convex optimization problem defining \( \hat{f}_{\text{erm}} \) by a convex one and derive from (approximate) solutions of this problem a classifier with bounded excess risk.

3.1 Construction of convex proxy

Recall the basic definition of convexity.

**Definition 32.** A set \( F \) is convex if, for any \( f \) and \( g \) in \( F \) and any \( t \in (0, 1) \),
\[ tf + (1 - t)g \in F. \]
A function \( h \) is convex on a convex domain \( C \) if, for any \( x \) and \( y \) in \( C \) and any \( t \in (0, 1) \),
\[ h(tx + (1 - t)y) \leq th(x) + (1 - t)h(y). \]

The following chapter will present several algorithms to solve approximately convex optimization problems where one has to minimize a convex function \( \Theta \) over a convex set \( C \). Typically sets of classifiers are not convex and, even if \( F \) is replaced by a convex sets \( \mathcal{F} \), the function that should be minimized to define \( \hat{f}_{\text{erm}} \),
\[ \theta_N(f) = P_N 1_{\{Y \neq f(X)\}}, \]
is not convex. Therefore, convex optimization algorithms cannot be used to approximate ERM \( \hat{f}_{\text{erm}} \). The idea of convex relaxation is to replace both \( F \) by a convex set \( \mathcal{F} \) and \( \theta_N \) by a convex function \( \Theta_N \) is such a way that one can derive from a minimizer of \( \Theta_N \) over \( \mathcal{F} \) a classifier with bounded
excess risk. Building these convex problems requires three main steps. The first step, sometimes called spinning, consists in replacing $Y \in \{0, 1\}$ by $Y^{(s)} = 2Y - 1 \in \{-1, 1\}$. The spinning variables satisfy

$$1_{\{Y \neq f(X)\}} = 1_{\{-Y^{(s)} f(X) > 0\}}.$$

The second step is to replace the class $F$ of classifiers by a convex class of functions called “soft classifiers”.

**Definition 33.** A soft classifier is a function $f^{(s)} : \mathcal{X} \rightarrow [-1, 1]$. The “hard” classifier $f^{(h)}$ associated with the soft classifier $f^{(s)}$ is $f^{(h)} = (1 + \text{Sign}(f^{(s)}))/2$.

Let $\mathcal{F}$ denote a convex set of soft classifiers. Typically, sets $\mathcal{F}$ of interest will be sets of linear functions $\mathcal{F} = \{a^T : a \in \mathcal{C}\}$, where $\mathcal{C}$ denotes a convex subset of $\mathbb{R}^d$. It may be that inputs $X_1, \ldots, X_N$ do not belong to $\mathbb{R}^d$, in this case, we usually use a set of functions $f_1, \ldots, f_d$ from $\mathcal{X} \rightarrow \mathbb{R}$ and consider as inputs the vectors

$$X_i = \begin{bmatrix} f_1(X_i) \\ \vdots \\ f_d(X_i) \end{bmatrix} \in \mathbb{R}^d.$$

Classical examples of convex sets $\mathcal{C}$ here are $\ell_p$-balls

$$B_p = \{a \in \mathbb{R}^d : \|a\|_p \leq 1\}, \quad \text{where} \quad \|a\|_p = \left( \sum_{i=1}^d |a_i|^p \right)^{1/p},$$

if $p = +\infty$ the $\ell_\infty$-norm is defined as usual as $\|a\|_\infty = \max_{1 \leq i \leq d} |a_i|$. One speaks about $\ell_p$-aggregation. Another example of interest is when $\mathcal{C}$ is the simplex

$$\Delta_d = \{a \in \mathbb{R}^d : \forall i \in \{1, \ldots, d\}, a_i \geq 0, \sum_{i=1}^d a_i = 1\},$$

one refers to this problem as convex aggregation problem.

The third step is to replace the non-convex loss function $1_{\{-Y^{(s)} f^{(s)}(X) > 0\}}$ by a convex surrogate. Let $\varphi$ denote a convex upper bound of the function $1_{\{x > 0\}}$. For example, $\varphi$ can denote the hinge loss $\varphi(x) = \max(1 + x, 0)$ or the logistic loss $\varphi(x) = \log_2(1 + e^x)$.

Given a set $\mathcal{F}$ of soft classifiers and a convex surrogate $\varphi$, the $\varphi$-ERM is defined by

$$\hat{f}_\varphi \in \arg\min_{f^{(s)} \in \mathcal{F}} P_N \ell_{\varphi, f^{(s)}}, \quad \text{where} \quad \ell_{\varphi, f^{(s)}}(z) = \varphi(-y^{(s)} f^{(s)}(x)) \quad .$$
3.2 Link with hard classifiers

Define the \( \varphi \)-bayes estimator

\[
    f^*_\varphi \in \arg\min_{f^{(s)}} P_{\ell_{\varphi,f^{(s)}}},
\]

where the minimum is taken among all functions \( f^{(s)} : \mathcal{X} \to [-1, 1] \). For any \( x \in \mathcal{X} \), it is clear that

\[
    f^*_\varphi(x) \in \arg\min_{\alpha \in \mathbb{R}} \mathbb{E}[\varphi(-Y^{(s)}\alpha)|X = x].
\]

**Theorem 34.** If \( \varphi \) is differentiable, then, the hard classifier associated to the \( \varphi \)-bayes estimators \( f^{(h)}_\varphi = (1 + \text{Sign}(f^*_\varphi))/2 \) is the bayes estimator

\[
    f^{(h)}_\varphi = \begin{cases} 
    1 & f^*_\varphi > 1/2 \\
    0 & f^*_\varphi \leq 1/2 
    \end{cases}
\]

**Proof.** The \( \varphi \)-loss function \( \ell_{\varphi,f^{(s)}}(z) \) is equal to \( \varphi(-f^{(s)}(x)) \) if \( y^{(s)} = 1 \), i.e. when \( y = 1 \) and to \( \varphi(f^{(s)}(x)) \) if \( y = 0 \), thus

\[
    \mathbb{E}[\ell_{\varphi,f^{(s)}}(X,Y)|X] = \mathbb{E}[Y \varphi(-f^{(s)}(X)) + (1 - Y)\varphi(f^{(s)}(X))|X] \\
    = \eta(X)\varphi(-f^{(s)}(X)) + (1 - \eta(X))\varphi(f^{(s)}(X)) \\
    = H_\eta(f^{(s)}(X)),
\]

where, for any \( \eta \in (0, 1) \) and \( \alpha \in \mathbb{R} \), \( H_\eta(\alpha) = \eta\varphi(-\alpha) + (1 - \eta)\varphi(\alpha) \), so

\[
    \forall x \in \mathcal{X}, \quad f^*_\varphi(x) \in \arg\min_{\alpha \in \mathbb{R}} H_\eta(f^{(s)}(X)).
\]

As \( H_\eta \) is differentiable, the minimum is achieved when the derivative is null. As \( H'_\eta(\alpha) = -\eta\varphi'(-\alpha) + (1 - \eta)\varphi'(\alpha) \), the condition \( H'_\eta(f^*_\varphi(x)) = 0 \) is equivalent to

\[
    \frac{\eta(x)}{1 - \eta(x)} = \frac{\varphi'(f^*_\varphi(x))}{\varphi'(-f^*_\varphi(x))}. \tag{3.1}
\]

As \( \varphi \) is convex, \( \varphi' \) is non-decreasing. As \( \eta(x) > 1/2 \) is equivalent to \( \varphi'(f^*_\varphi(x)) > \varphi'(-f^*_\varphi(x)) \) by (3.1), it is also equivalent to \( f^*_\varphi(x) > 0 \). \( \square \)

Perhaps an even more interesting link between the classification problem and its convex relaxation is provided by the following result, which links the excess risk of the hard classifier \( f^{(h)} \) associated to a soft classifier \( f^{(s)} \) with the \( \varphi \)-excess risk of \( f^{(s)} \).

**Lemma 35** (Zhang’s lemma). Let \( \varphi \) denote a non-decreasing, convex, non-negative function such that \( \varphi(0) = 1 \). Denote by \( \tau(\eta) = \inf_{\alpha \in \mathbb{R}} H_\eta(\alpha) \) and assume that there exist constants \( c > 0 \) and \( \gamma \in [0, 1] \) such that

\[
    \forall \eta \in [0, 1], \quad |\eta - 1/2| \leq c(1 - \tau(\eta))^{\gamma}. \tag{3.2}
\]
Then, for any soft classifier \( f^{(s)} \), the associated hard classifier \( f^{(h)} = (1 + \text{Sign}(f^{(s)}))/2 \) satisfies
\[
E(f^{(h)}) \leq 2eE_{\varphi}(f^{(s)})^\gamma, \quad E_{\varphi}(f^{(s)}) = P[\ell_{\varphi,f^{(s)}}] - \min_f P[\ell_{\varphi,f}].
\]

Before proceeding with the proof of Zhang’s lemma, it is interesting to check the condition (3.2) appearing in this lemma on convex surrogates of interests. Start with the hinge loss \( \varphi(x) = \max(1 + x; 0) \). In this case,
\[
H(\gamma) = \max(1 + \gamma) + (1 - \gamma) \max(1 + \gamma, 0)
\]
\[
= \begin{cases} 
\gamma(1 - \alpha) & \text{if } \alpha \leq -1 \\
\gamma(1 - \alpha) + (1 - \gamma)(1 + \alpha) = 1 + (1 - 2\gamma)\alpha & \text{if } -1 < \alpha < 1 \\
(1 - \gamma)(1 + \alpha) & \text{if } \alpha \geq 1
\end{cases}
\]
It follows that \( \tau(\eta) = 2 \min(\eta, (1 - \eta)) \). Hence, if \( \eta < 1/2 \),
\[
|\eta - 1/2| = 1/2 - \eta = \frac{1}{2}(1 - 2 \min(\eta, (1 - \eta))) = \frac{1}{2}(1 - \tau(\eta)).
\]
If \( \eta > 1/2 \),
\[
|\eta - 1/2| = \eta - 1/2 = \frac{1}{2}(1 - 2(1 - \eta)) = \frac{1}{2}(1 - 2 \min(\eta, (1 - \eta))) = \frac{1}{2}(1 - \tau(\eta)).
\]
Therefore, Condition (3.2) of Zhang’s Lemma holds with \( c = 1 \).

Consider now the logistic loss \( \varphi(x) = \log(1 + e^x) \). Then
\[
H(\eta)(\alpha) = \eta \log(1 + e^{-\alpha}) + (1 - \eta) \log(1 + e^\alpha).
\]
Deriving this expression yields
\[
H'(\eta)(\alpha) = \frac{1}{\log 2} \frac{-\eta + (1 - \eta)e^\alpha}{e^\alpha + 1}.
\]
It follows that \( H(\eta) \) achieves its minimum at \( \alpha = \log(\eta/(1 - \eta)) \). This minimum is equal to
\[
\tau(\eta) = H(\eta)\left(\log\left(\frac{\eta}{1 - \eta}\right)\right) = \eta \log_2 \left(1 + \frac{1 - \eta}{\eta}\right) + (1 - \eta) \log_2 \left(1 + \frac{\eta}{1 - \eta}\right)
\]
\[
= -\eta \log_2(\eta) - (1 - \eta) \log_2(1 - \eta).
\]
It follows that
\[
1 - \tau(\eta) = \frac{1}{\log 2} \left(-\log(1/2) + \eta \log(\eta) + (1 - \eta) \log(1 - \eta)\right)
\]
\[
= \frac{1}{\log 2} \left(\eta \log\left(\frac{\eta}{1/2}\right) + (1 - \eta) \log\left(\frac{1 - \eta}{1/2}\right)\right). \quad (3.3)
\]
At this point, the following result is useful.
Lemma 36 (Pinsker’s inequality). Let \( P, Q \) denote two probability distributions. Denote by \( \mu \) a measure dominating both \( P \) and \( Q \) and let \( p, q \) denote the densities of \( P \) and \( Q \) respectively with respect to \( \mu \). Then

\[
\frac{1}{2} \left( \int |p - q| \, d\mu \right)^2 \leq \int p \log \left( \frac{p}{q} \right) \, d\mu .
\]

Proof of Pinsker’s inequality. Let

\[
\psi(x) = x \log(x) - x + 1 .
\]

Then, \( \psi(0) = 1, \psi(1) = 0, \psi'(1) = 0, \psi''(x) = 1/x \geq 0 \), so \( \psi(x) \geq 0 \) for any \( x > 0 \). Let also \( g(x) = (x - 1)^2 - (4/3 + 2x/3)\psi(x) \). As \( g(1) = g'(1) = 0 \) and \( g''(x) = -4\psi(x)/3 \leq 0 \), it holds that, for any \( x > 0 \), there exists \( \xi \) such that \( |\xi - 1| < |x - 1| \) and

\[
g(x) = g(1) + g'(1)(x - 1) + g''(\xi)(x - 1)^2/2 \leq 0 .
\]

Therefore

\[
\forall x > 0, \quad (x - 1)^2 \leq \left( \frac{4}{3} + \frac{2x}{3} \right) \psi(x) .
\]

It follows that

\[
\int |p - q| \, d\mu = \int q \left| \frac{p}{q} - 1 \right| \, d\mu
\leq \int q \sqrt{\left( \frac{4}{3} + \frac{2p}{3q} \right) \psi \left( \frac{p}{q} \right)} \, d\mu
\leq \sqrt{\int \left( \frac{4q}{3} + \frac{2p}{3} \right) \, d\mu} \sqrt{\int q \psi \left( \frac{p}{q} \right) \, d\mu}
= \sqrt{2} \sqrt{\int \left( p \log \left( \frac{p}{q} \right) - p + q \right) \, d\mu}
= \sqrt{2} \sqrt{\int p \log \left( \frac{p}{q} \right) \, d\mu} .
\]

\( \square \)

Going back to (58), applying Pinsker’s inequality with the Bernoulli distributions

\[
P = B(\eta), \quad Q = B(1/2) ,
\]

gives
\[2(\eta - 1/2)^2 = \frac{1}{2} \left( |\eta - 1/2| + |(1 - \eta) - 1/2| \right)^2 \leq \eta \log \left( \frac{\eta}{1/2} \right) + (1 - \eta) \log \left( \frac{1 - \eta}{1/2} \right).\]

Therefore, \(1 - \tau(\eta) \geq (2/\log 2)(\eta - 1/2)^2\). Hence, Zhang’s condition (3.2) holds with \(\gamma = 1/2\) and \(c = \sqrt{\log(2)/2}\).

**Proof of Zhang’s lemma.** By the representation of the excess risk theorem
\[\mathcal{E}(f^{(h)}) = \mathbb{E}||2\eta(X) - 1|1_{\{f^{(s)}(X) \neq f^*(X)\}}||.\]

By definition of the hard classifier \(f^{(h)}\),
\[1_{\{f^{(h)}(X) \neq f^*(X)\}} = 1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}};\]

so
\[\mathcal{E}(f^{(h)}) = \mathbb{E}[|2\eta(X) - 1|1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}}].\]

By Zhang’s condition (3.2),
\[\mathcal{E}(f^{(h)}) \leq 2c\mathbb{E}[(1 - \tau(\eta(X)))^{\gamma}1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}}].\]

By Jensen’s inequality,
\[\mathcal{E}(f^{(h)}) \leq 2c\mathbb{E}[(1 - \tau(\eta(X)))^{\gamma}1_{\{f^{(s)}(X)(\eta(X) - 1/2) < 0\}}]^{\gamma}. \tag{3.4}\]

By convexity of \(\varphi\),
\[H_{\eta(x)}(f^{(s)}(x)) = \eta(x)\varphi(-f^{(s)}(x)) + (1 - \eta(x))\varphi(f^{(s)}(x)) \geq \varphi((1 - 2\eta(x))f^{(s)}(x)).\]

Therefore, if \(f^{(s)}(x)(\eta(x) - 1/2) < 0\), as \(\varphi\) is non-decreasing
\[H_{\eta(x)}(f^{(s)}(x)) \geq \varphi(0) = 1.\]

Hence, if \(f^{(s)}(x)(\eta(x) - 1/2) < 0\),
\[(1 - \tau(\eta(x)))1_{\{f^{(s)}(x)(\eta(x) - 1/2) < 0\}} = 1 - \tau(\eta(x)) \leq H_{\eta(x)}(f^{(s)}(x)) - \mathbb{E}[\ell_{\varphi,f^*_s}(X,Y)|X = x]\]
\[= \mathbb{E}[\ell_{\varphi,f^{(s)}}(X,Y)|X = x] .\]
Moreover, if \( f^{(s)}(x)(\eta(x) - 1/2) \geq 0 \), then
\[
(1 - \tau(\eta(x)))I_{\{f^{(s)}(\eta(x) - 1/2) < 0\}} = 0 \leq H_{\eta(x)}(f^{(s)}(x)) - \tau(\eta(x))
= \mathbb{E}[\ell_{\phi,f^{(s)}}(X,Y) | X = x] .
\]

It follows that, for all \( x \in X \),
\[
(1 - \tau(\eta(x)))I_{\{f^{(s)}(\eta(x) - 1/2) < 0\}} \leq \mathbb{E}[\ell_{\phi,f^{(s)}}(X,Y) | X = x] .
\]

Integrating over \( x \in X \) shows then that
\[
\mathbb{E}[(1 - \tau(\eta(X)))I_{\{f^{(s)}(\eta(X) - 1/2) < 0\}}] \leq \mathcal{E}_{\ell_{\phi}}(f^{(s)}) .
\]

Plugging this inequality into (3.4) concludes the proof. \( \square \)

### 3.3 Bounding the \( \phi \)-excess risk

Thanks to Zhang’s lemma, one can bound the excess risk of hard classifiers by bounding \( \phi \)-excess risks of soft classifiers. Recall that from chapters 1 and 2, by Vapnik’s lemma
\[
\mathbb{E}[\ell_{\phi,f^{(s)}}(X,Y)] \leq \inf_{f \in \mathcal{F}} \mathbb{E}[\ell_{\phi,f}(X,Y)] + 2 \sup_{f^{(s)} \in \mathcal{F}} \mathbb{E}[(P_N - P)\ell_{\phi,f^{(s)}}] + \sqrt{\frac{2\log(1/\delta)}{N}} .
\]

Then, by the bounded difference inequality, \( \ell_{\phi,f^{(s)}}, \) taking values in \([-1,1]\), for any \( \delta \in (0,1) \), with probability at least \( 1 - \delta \),
\[
\mathbb{E}[\sup_{f^{(s)} \in \mathcal{F}} \| (P_N - P)\ell_{\phi,f^{(s)}} \|] \leq \mathbb{E}[\sup_{f^{(s)} \in \mathcal{F}} \| (P_N - P)\ell_{\phi,f^{(s)}} \|] + \sqrt{\frac{2\log(1/\delta)}{N}} .
\]

Then, by the symmetrization lemma
\[
\mathbb{E}[\sup_{f^{(s)} \in \mathcal{F}} \| (P_N - P)\ell_{\phi,f^{(s)}} \|] \leq 2R_N(\phi, \mathcal{F}) ,
\]
where
\[
R_N(\phi, \mathcal{F}) = \sup_{z \in 2^N} \mathbb{E} \left[ \sup_{f^{(s)} \in \mathcal{F}} \left\| \frac{1}{N} \sum_{i=1}^N \epsilon_i \phi(-y_i f^{(s)}(x_i)) \right\| \right] .
\]

At this point, it is useful to invoke the contraction inequality of Ledoux and Talagrand: If \( \phi \) is \( L \)-Lipschitz and satisfies \( \phi(0) = 0 \), then for any set \( B \subset \mathbb{R}^N \),
\[
\mathbb{E} \left[ \sup_{b \in B} \left\| \frac{1}{N} \sum_{i=1}^N \epsilon_i \phi(b_i) \right\| \right] \leq 2L \mathbb{E} \left[ \sup_{b \in B} \left\| \frac{1}{N} \sum_{i=1}^N \epsilon_i b_i \right\| \right]
\]
Applying this inequality to \( \phi(\cdot) = \varphi(\cdot) - 1 \) and assuming that \( \varphi \) is \( L \)-Lipschitz as are both the hinge loss and the logistic loss, it holds \( R_N(\varphi, \mathcal{F}) \leq 2LR_N(\mathcal{F}) \), where

\[
R_N(\mathcal{F}) = \sup_{z \in \mathcal{Z}} \mathbb{E} \left[ \sup_{f(\cdot) \in \mathcal{F}} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i y_i f^{(s)}(x_i) \right| \right].
\]

Ultimately, the hard thresholded estimator \( \hat{f} = (1 + \text{Sign}(\hat{f}_\varphi))/2 \) satisfies

\[
\mathcal{E}(\hat{f}) \leq 2c \left( \inf_{f \in \mathcal{F}} \{ \mathcal{E}_{\varphi}(f) \} + 8LR_N(\mathcal{F}) + 2 \sqrt{2 \log(1/\delta)/N} \right)^\gamma. \tag{3.5}
\]

As in Chapter 2, the tricky part is to bound from above the Rademacher complexity \( R_N(\mathcal{F}) \) in (3.5). Chapter 2 developed general approaches for this problem. The following sections provide other computations for widely used classes of soft classifiers.

### 3.4 Boosting

Boosting is an example of linear soft classifiers indexed by the \( \ell_1 \)-ball \( \mathcal{B}_1 \). Let \( f_1, \ldots, f_d \) denote a collection of hard classifiers and let \( \mathcal{F} \) denotes the convex hull of the classifiers \( f_i, -f_i \):

\[
\mathcal{F} = \{ \sum_{i=1}^{M} a_i f_i : \sum_{i=1}^{M} a_i \leq 1 \}.
\]

In other words, defining the vector valued inputs

\[
X_i = \begin{bmatrix} f_1(X_i) \\ \vdots \\ f_d(X_i) \end{bmatrix}, \quad \forall x \in \mathcal{X}, \quad x = \begin{bmatrix} f_1(x) \\ \vdots \\ f_d(x) \end{bmatrix}
\]

and the convex set \( \mathcal{C} = \mathcal{B}_1, \mathcal{F} = \{ a^T \cdot a \in \mathcal{C} \} \).

**Theorem 37.** The Rademacher complexity of the boosting class \( \mathcal{F} \) satisfies

\[
R_N(\mathcal{F}) \leq \sqrt{\frac{2 \log(4d)}{N}}.
\]

In particular, if \( \varphi \) is either the hinge loss or the logistic loss and \( \hat{f} \) is the hard classifier associated to the \( \varphi \)-ERM \( \hat{f}_\varphi \) over \( \mathcal{F} \), it holds

\[
\mathcal{E}(\hat{f}) \leq 2c \left( \inf_{f \in \mathcal{F}} \{ \mathcal{E}_{\varphi}(f) \} + 8L \sqrt{\frac{2 \log(4d)}{N}} + 2 \sqrt{\frac{2 \log(1/\delta)}{N}} \right)^\gamma.
\]
Proof. By definition,

\[ R_N(\mathcal{F}) = \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \sup_{a \in B_1} \left| a^T \left( \frac{1}{N} \sum_{i=1}^{N} \epsilon_i y_i X_i \right) \right| \right] . \]

As, for any vectors \( a, b \) in \( \mathbb{R}^d \), it holds

\[ a^T b \leq \|a\|_1 \|b\|_\infty, \]

where \( B_z \) is the set of \( 2d \) vectors

\[ B_z = \left\{ \pm \begin{bmatrix} y_1 e_1^T X_1 \\ \vdots \\ y_N e_1^T X_N \end{bmatrix}, \ldots, \pm \begin{bmatrix} y_1 e_d^T X_1 \\ \vdots \\ y_N e_d^T X_N \end{bmatrix} \right\}. \]

It holds that

\[ R_N(\mathcal{F}) = \sup_{z \in \mathbb{Z}^N} \mathbb{E} \left[ \max_{b \in B_z} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i b_i \right| \right] . \]

The vectors \( b \in B_z \) have coordinates at most 1, so \( \|b\|^2 \leq N \), hence \( \Delta(B_z) \leq \sqrt{N} \), thus, by Lemma 18 it follows that

\[ R_N(\mathcal{F}) \leq \sqrt{\frac{2 \log(4d)}{N}} . \]

This proves the first part of the theorem, the second part comes from Eq (3.5).

\[ \square \]

3.5 Support Vector Machine

In this section \( \varphi \) denotes the hinge loss \( \varphi(x) = \max(1 + x, 0) \) and \( \mathcal{F} \) is the unit ball of a Hilbert space \( W \).

3.5.1 Reproducing kernel Hilbert space

Start with the definition of positive semi-definite kernels.

**Definition 38 (PSD kernels).** A function \( K : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \) is called a positive semi-definite kernel (PSD kernel) if, for any \( x_1, \ldots, x_N \in \mathcal{X} \), the matrix \( K \) such that \( K_{i,j} = K(x_i, x_j) \) is symmetric, positive semi-definite, that is if \( K^T = K \) and, for any \( a \in \mathbb{R}^N \), \( a^T K a \geq 0 \).
Example 1. The function $K(x, y) = x^T y$ is a PSD kernel on $\mathbb{R}^d$.

Example 2. The function $K(x, y) = e^{-\|x-y\|^2/(2\sigma^2)}$ is a PSD kernel on $\mathbb{R}^d$.

Definition 39. Let $W$ denote a Hilbert space of functions $f : \mathcal{X} \to \mathbb{R}$. A symmetric kernel $K$ is called reproducing kernel of $W$ if

(i) for any $x \in \mathcal{X}$, $K(x, \cdot) \in W$,

(ii) for any $x \in \mathcal{X}$ and any $f \in W$, $\langle f, K(x, \cdot) \rangle_W = f(x)$.

If $W$ admits a reproducing kernel $K$, it is called a reproducing kernel Hilbert space, with kernel $K$.

The following lemma makes a link between these notions.

Lemma 40. A reproducing kernel is positive semi-definite.

Proof. Let $K$ denote a reproducing kernel, $x \in \mathcal{X}^N$ and $K$ denote the associated matrix with entries $K_{i,j} = K(x_i, x_j)$. Let $a \in \mathbb{R}^N$. By the reproducing property

$$a^T K a = \left\langle \sum_{i=1}^N a_i K(x_i, \cdot), \sum_{i=1}^N a_i K(x_i, \cdot) \right\rangle_W = \left\| \sum_{i=1}^N a_i K(x_i, \cdot) \right\|_W^2 \geq 0 .$$

Example 3. If $\varphi_1, \ldots, \varphi_M$ denote orthonormal functions in $L^2(\mu)$, then $K(x, y) = \sum_{i=1}^M \varphi_i(x) \varphi_i(y)$ is a reproducing kernel in the space $W = \{ f = \sum_{i=1}^M a_i \varphi_i, \ a \in \mathbb{R}^M \}$ endowed with the $L^2(\mu)$ inner product.

It is sufficient to check that $K$ is a PSD kernel satisfying the reproducing property. First,

$$a^T K a = \sum_{j=1}^M \left( \sum_{i=1}^N a_i \varphi_j(x_i) \right)^2 \geq 0 .$$

Hence, $K$ is PSD. Second, let $f \in W$, so $f(x) = \sum_{i=1}^M a_i \varphi_i$. Thus

$$\langle f, K(x, \cdot) \rangle = \sum_{i=1}^M \sum_{j=1}^M a_i \varphi_j(x) \langle \varphi_i, \varphi_j \rangle_{L^2(\mu)} = \sum_{i=1}^M a_i \varphi_i(x) = f(x) .$$

Example 4. $K(x, y) = x^T y$ is a reproducing kernel of $W = \{ w^T \cdot, \ w \in \mathbb{R}^d \}$ equipped with the inner product $\langle w^T \cdot, (w')^T \cdot \rangle = w^T w'$.
3.5. SUPPORT VECTOR MACHINE

The sup-norm in the RKHS is bounded from above by the norm of the RKHS as shown by the following proposition.

**Proposition 41.** Let \( W \) denote a RKHS with PSD kernel \( K \) such that \( \sup_{x \in X} K(x, x) = k_\infty < \infty \). Then,

\[
\forall f \in W, \quad \sup_{x \in X} |f(x)| \leq \|f\|_W \sqrt{k_\infty}.
\]

**Proof.** Use the reproducing property to write

\[
|f(x)| = \langle f, K(x, \cdot) \rangle_W.
\]

Then by Cauchy-Schwarz inequality,

\[
|f(x)| \leq \|f\|_W \sqrt{\langle K(x, \cdot), K(x, \cdot) \rangle_W}.
\]

By the reproducing property,

\[
\langle K(x, \cdot), K(x, \cdot) \rangle_W = K(x, x) \leq k_\infty.
\]

\[ \square \]

3.5.2 Representer theorem

The key to compute the \( \varphi \)-ERM \( \hat{f}_{\varphi} \) over the unit ball of the possibly infinite dimensional space \( W \) is that it actually takes value in a finite dimensional space thanks to the following result.

**Theorem 42** (Representer theorem). Let \( W \) denote a RKHS with PSD \( K \) and let \( G : \mathbb{R}^N \rightarrow \mathbb{R} \) denote any function. Let \( x_1, \ldots, x_N \in X \). For any \( r > 0 \), let \( B_W(r) = \{ f \in W : \|f\|_W \leq r \} \), let \( W_N = \{ a = \sum_{i=1}^N a_i K(x_i, \cdot), \ a \in \mathbb{R}^N \} \) and \( B_{W_N}(r) = \{ f \in W_N : \|f\|_W \leq r \} \). Then

\[
\min_{f \in B_W(r)} G(f(x_1), \ldots, f(x_N)) = \min_{f \in B_{W_N}(r)} G(f(x_1), \ldots, f(x_N))
\]

\[
= \min_{a \in \mathbb{R}^N : a^T a \leq r^2} G(f_a(x_1), \ldots, f_a(x_N)).
\]

**Proof.** Let \( f \in W \), and write \( f = f_N + f_N^\perp \) with \( f_N \in W_N \) and \( f_N^\perp \in W_N^\perp \). Since \( K(x_i, \cdot) \in W_N \), it holds \( f^T(x_i) = \langle f^T, K(x_i, \cdot) \rangle_W = 0 \) so \( f(x_i) = f_N(x_i) \). Moreover, by Pythagoras theorem \( \|f_N\|_W \leq \|f\|_W \). Hence, for any \( f \in B_W(r) \)

\[
G(f(x_1), \ldots, f(x_N)) = G(f_N(x_1), \ldots, f_N(x_N)),
\]
with $f_N \in W_N$ and $\|f_N\|_W \leq r$, so

$$
\min_{f \in B_w(r)} G(f(x_1), \ldots, f(x_N)) \geq \min_{f \in B_{w_N}(r)} G(f(x_1), \ldots, f(x_N)) .
$$

Conversely, $B_{w_N}(r) \subset B_w(r)$ so

$$
\min_{f \in B_w(r)} G(f(x_1), \ldots, f(x_N)) \leq \min_{f \in B_{w_N}(r)} G(f(x_1), \ldots, f(x_N)) .
$$

This shows the first part of the theorem. The second is a direct consequence of the reproducing property, since

$$
\|f_a\|_W^2 = \sum_{1 \leq i, j \leq N} a_i a_j \langle K(x_i, \cdot), K(x_j, \cdot) \rangle_W = a^TKa .
$$

3.5.3 Excess risk of $\phi$-ERM

Remark that SVM algorithm can be recast as a $\ell_2$-aggregation problem. Consider as input variables

$$
\forall i \in \{1, \ldots, N\}, \quad X_i = \begin{bmatrix}
K(X_1, X_i) \\
\vdots \\
K(X_N, X_i)
\end{bmatrix}
$$

and, as input space the set of vectors $x \in \mathbb{R}^d$ such that

$$
\forall x \in \mathcal{X}, \quad x = \begin{bmatrix}
K(X_1, x) \\
\vdots \\
K(X_N, x)
\end{bmatrix}
$$

Define also

$$
B_w(r) = \{ f \in W : \|f\|_W \leq r \}, \\
B_{w_N}(r) = \{ f \in W_N : \|f\|_W \leq r \} = \{ a^T, \ a^TKa \leq r^2 \} .
$$

By the representer theorem, it holds that

$$
\arg\min_{f \in B_w(r)} P_N \ell_\phi, f(X_i, Y_i) = \arg\min_{f \in B_{w_N}(r)} P_N \ell_\phi, f(X_i, Y_i)
$$

The following theorem computes the Rademacher complexity of balls in RKHS. from which an excess risk bound for the hard classifier associated with the $\varphi$-ERM over this ball is obtained.
3.5. SUPPORT VECTOR MACHINE

Theorem 43. Let $W$ denote a RKHS with PSD $K$ such that $k_\infty < \infty$ and let $\mathcal{F} = B_W(r)$. Then

$$R_N(\mathcal{F}) \leq r \sqrt{\frac{\text{Tr}(K)}{N}} \leq r \sqrt{\frac{k_\infty}{N}}.$$ 

Denote by $\hat{f} = (1 + \text{Sign}(\hat{f}_\varphi))/2$ the hard classifier associated with the $\varphi$-ERM

$$\hat{f}_\varphi \in \arg\min_{f \in B_W(r)} P_N \ell_{\varphi, f}.$$ 

For any $\delta \in (0, 1)$, with probability $1 - \delta$,

$$\mathcal{E}(\hat{f}) \leq \inf_{f \in \mathcal{F}} \{\mathcal{E}_\varphi(f)\} + 8\sqrt{\frac{k_\infty}{N}} + 2\sqrt{2 \log(1/\delta)}.$$ 

Proof. By Cauchy-Schwarz inequality,

$$R_N(\mathcal{F}) = \sup_{z \in Z^N} \mathbb{E} \left[ \sup_{f \in B_W(r)} \left\langle \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot), f \right\rangle_W \right] \leq r \sup_{z \in Z^N} \mathbb{E}\left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot) \right\|^2_W \right] \leq r \sup_{z \in Z^N} \sqrt{\mathbb{E}\left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot) \right\|^2_W \right]}.$$ 

Moreover, by the representing property

$$\mathbb{E}\left[ \left\| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i K(x_i, \cdot) \right\|^2_W \right] = \frac{1}{N^2} \sum_{1 \leq i, j \leq N} K(x_i, x_j) \mathbb{E}[\epsilon_i \epsilon_j] = \frac{\text{Tr}(K)}{N^2} \leq \frac{k_\infty}{N}.$$ 

This shows the first part of the theorem. The second part comes from (S8) combined with the values of $c$ and $\gamma$ obtained for the hinge loss right after Zhang’s lemma. \qed
Chapter 4
Convex optimization

This chapter presents various algorithms to solve convex problems. In machine learning, these algorithms are useful to solve the following problems:

$$\min_{f \in \mathcal{F}} \Theta_N(f), \quad \min_{f \in \mathcal{F}} \Theta(f)$$

where $\mathcal{F}$ is a convex set of soft classifiers and, for a convex surrogate $\varphi$

$$\Theta_N(f) = P_N \ell_\varphi = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y_i f(X_i)), \quad \Theta(f) = P\ell_\varphi = \mathbb{E}[\varphi(-Y f(X))]$$

There is an important difference between these problems for machine learning. In the first case, convex (deterministic) optimization algorithms are used to obtain an approximation $\hat{f}_\varphi^{(\epsilon)}$ of the ERM $\hat{f}_\varphi$ with a deterministic control of the approximation error by $\epsilon > 0$. The risk of this $\epsilon$-ERM, $\hat{f}_\varphi^{(\epsilon)}$ estimator is then bounded from above by the sum of $\epsilon$ (that depends on the number of iterations of the algorithm) and the estimation error of the ERM (that has been bounded from above in the previous chapters). In the second case, stochastic convex approximation algorithms directly produce estimators $\hat{f}$. The approximation error of these algorithm is the $\varphi$-excess risk of $\hat{f}$. Therefore, stochastic convex optimization does not rely on ERM theory. This error will be bounded in expectation only and not with high probability. There is another important difference between these approaches, while deterministic optimization requires an access to all data simultaneously, these can be made available on line for stochastic approximation, which considerably saves memory resources when handling very large datasets.
CHAPTER 4. CONVEX OPTIMIZATION

4.1 Convex problems

Definition 44 (Convex problems). An optimization problem of the form
\[ \min_{x \in C} \theta(x) \] is called a convex problem if \( C \) is a convex set and \( \theta \) is a convex function.

Assume that the objective function \( \theta \) is defined on a domain \( D \subset \mathbb{R}^d \) and that there exists a vector \( c \) such that, for all \( x \in D \), \( \theta(x) = c^T x \). In this case, even if \( D \) is not a convex set, the problem \( \min_{x \in D} \theta(x) \) is equivalent to a convex problem. To see why, introduce the following definition.

Definition 45 (Convex hull). Let \( D \subset \mathbb{R}^d \). The convex hull of \( D \), \( H(D) \), is the set of convex combinations of elements of \( D \):
\[
H(D) = \{ \sum_{i=1}^{N} a_i x_i, \; x_i \in D, \; a \in \Delta_N \} .
\]

The convex hull of any set \( D \subset \mathbb{R}^d \) is the smallest convex containing \( D \) and the following result holds.

Proposition 46. For any \( c \in \mathbb{R}^d \) and any non-empty subset \( D \subset \mathbb{R}^d \),
\[
\min_{x \in D} c^T x = \min_{x \in H(D)} c^T x .
\]

Proof. Since \( D \subset H(D) \), it is clear that
\[
\min_{x \in D} c^T x \geq \min_{x \in H(D)} c^T x .
\]

Hence, it is sufficient to prove the reverse inequality. Let \( x \in H(D) \), by definition, there exist \( x_1, \ldots, x_N \) in \( D \) and \( a \in \Delta_N \) such that \( x = \sum_{i=1}^{N} a_i x_i \). Thus
\[
c^T x = \sum_{i=1}^{N} a_i c^T x_i .
\]

Now, all \( c^T x_i \geq \min_{x \in D} c^T x \). As \( a \in \Delta_N \), all \( a_i \geq 0 \) and \( \sum_{i=1}^{N} a_i = 1 \), so
\[
c^T x \geq \sum_{i=1}^{N} a_i \min_{x \in D} c^T x = \min_{x \in D} c^T x .
\]

As this holds for all \( x \in H(D) \), it follows that
\[
\min_{x \in H(D)} c^T x \geq \min_{x \in D} c^T x .
\]

This concludes the proof of the second inequality, hence, the proof of the proposition. \( \square \)
A fundamental tool in this chapter is the notion of sub-gradient.

**Definition 47** (sub-gradient). Let $D \subset \mathbb{R}^d$ and let $\theta : D \to \mathbb{R}$. A vector $g \in \mathbb{R}^d$ is called a sub-gradient of $\theta$ at $x \in D$ if

$$\forall y \in D, \quad \theta(x) - \theta(y) \leq g^T(x - y).$$

The set of sub-gradients of $\theta$ at $x$ is called the sub-differential of $\theta$ at $x$ and is denoted by $\partial \theta(x)$.

Sub-gradients play the roles of gradients but, contrary to gradients, they are well defined at any point $x$ if $\theta$ is convex as shown by the following (admitted) result.

**Theorem 48.** If $\theta : C \to \mathbb{R}$ is convex, then, for any $x \in C$, $\partial \theta(x) \neq \emptyset$. Moreover, if $\theta$ is differentiable at $x \in C$, then $\partial \theta(x) = \{\nabla \theta(x)\}$.

The following result shows that, for convex functions, it is sufficient to show that 0 is a sub-gradient of $\theta$ at $x$ to prove that $x$ is a minimizer of $\theta$.

**Theorem 49.** Let $\theta$ denote a convex function over a convex domain $C$. The following are equivalent.

(i) $x$ is a global minimum of $\theta$.

(ii) $x$ is a local maximum of $\theta$.

(iii) $0 \in \partial \theta(x)$.

**Proof.** By definition of sub-differentials, $0 \in \partial \theta(x)$ if and only if

$$\forall y \in C, \quad \theta(x) - \theta(y) \leq 0,$$

that is if and only if $x$ is a global minimum of $\theta$.

Moreover, if $x$ is a local minimum of $\theta$, there exists $r > 0$ such that, $\forall y \in C$ satisfying $\|y - x\| \leq r$, $\theta(x) \leq \theta(y)$. Let $y' \in C$, there exists $\epsilon \in (0, 1)$ such that $y = x + \epsilon(y' - x) \in C$ and satisfies $\|y - x\| \leq r$. Therefore

$$\theta(y) \geq \theta(x).$$

By convexity of $\theta$,

$$\theta(y) = \theta(\epsilon y' + (1 - \epsilon)x) \leq \epsilon \theta(y') + (1 - \epsilon) \theta(x).$$

Thus

$$\epsilon \theta(y') + (1 - \epsilon) \theta(x) \geq \theta(x),$$

which is equivalent to

$$\theta(y') \geq \theta(x).$$

As this holds for any $y' \in C$, this shows that $x$ is a global minimizer of $\theta$. \qed
4.2 Gradient descent

Informally, if \( \|y - x\| \) is small then \( \theta(y) \) is close to its linear approximation

\[
\theta(y) \approx \theta(x) + \nabla \theta(x)^T (y - x).
\]

Therefore, if, for a small \( \eta > 0 \), \( y = x - \eta \nabla \theta(x) \)

\[
\theta(y) \approx \theta(x) - \eta \| \nabla \theta(x) \|^2 \leq \theta(x).
\]

Gradient descent exploits this heuristic, building recursively a sequence of approximate minimizers of \( \theta \) as follows.

<table>
<thead>
<tr>
<th>input</th>
<th>: ( x_1 \in \mathcal{C} ): initial point, ((\eta_t)_t): step sizes, ( T ): stopping time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>for ( t = 1 ) to ( t = T ), do</td>
</tr>
<tr>
<td>2</td>
<td>find ( g_t \in \partial \theta(x_t) ),</td>
</tr>
<tr>
<td>3</td>
<td>( x_{t+1} = x_t - \eta_t g_t ).</td>
</tr>
<tr>
<td>4</td>
<td>end</td>
</tr>
<tr>
<td>5</td>
<td>Return ( x_T = T^{-1} \sum_{t=1}^{T} x_t ) or ( x_T \in \arg\min_{x \in {x_1, \ldots, x_T}} \theta(x) ).</td>
</tr>
</tbody>
</table>

**Algorithm 1**: Gradient descent (GD).

The following result analyses performance of gradient descent algorithm.

**Theorem 50.** Assume that \( \theta \) is convex, \( L \)-Lipschitz on \( \mathbb{R}^d \) and denote by \( x^* \in \arg\min_{x \in \mathbb{R}^d} \theta(x) \). If \( \|x_1 - x^*\| \leq R^2 \) and, for any \( t \geq 1 \), \( \eta_t = \eta = R/(L\sqrt{T}) \), then, the outputs of GD satisfy

\[
\theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}, \quad \theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}.
\]

**Proof.** Let \( t \in \{1, \ldots, T\} \), then, by definition of a sub-gradient,

\[
\theta(x_t) - \theta(x^*) \leq g_t^T (x_t - x^*) .
\]

By definition of the gradient descent algorithm, \( g_t = \eta_t^{-1} (x_{t+1} - x_t) \), so

\[
\theta(x_t) - \theta(x^*) \leq \frac{1}{\eta_t} (x_{t+1} - x_t)^T (x_t - x^*) .
\]

Using the equality \( a^T b = (\|a\|^2 + \|b\|^2 - \|a - b\|^2)/2 \), it follows that

\[
\theta(x_t) - \theta(x^*) \leq \frac{1}{2\eta_t} (\|x_{t+1} - x_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) .
\]
4.2. GRADIENT DESCENT

By definition of the gradient descent algorithm, it follows that

$$\theta(x_t) - \theta(x^*) \leq \frac{1}{2\eta_t} (\eta_t^2 \|g_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) .$$

As $\theta$ is $L$-Lipschitz, the sub-gradient $g_t$ has $L^2$-norms bounded from above by $L$, thus

$$\eta_t (\theta(x_t) - \theta(x^*)) \leq \frac{1}{2} (\eta_t^2 L^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2) .$$  \hspace{1cm} (4.1)

Summing up these inequalities and dividing by $\sum_{t=T_0}^T \eta_t$ yields

$$\sum_{t=T_0}^T \eta_t (\theta(x_t) - \theta(x^*)) \leq \frac{1}{2} \left( L^2 \sum_{t=T_0}^T \frac{\eta_t^2}{\eta_t} + \|x_1 - x^*\|^2 \right) .$$

By convexity of $\theta$, this yields

$$\theta \left( \frac{\sum_{t=T_0}^T \eta_t x_t}{\sum_{t=T_0}^T \eta_t} \right) - \theta(x^*) \leq \frac{1}{2} \left( L^2 \frac{\sum_{t=T_0}^T \eta_t^2}{\sum_{t=T_0}^T \eta_t} + \frac{R^2}{\sum_{t=T_0}^T \eta_t} \right) .$$  \hspace{1cm} (4.2)

Taking $T_0 = 1$ and $\eta_t = \eta$ as in the theorem yields

$$\theta(x_t) - \theta(x^*) \leq \frac{1}{2} \left( L^2 \eta + \frac{R^2}{\eta T} \right) .$$

With the choice of $\eta$, this gives

$$\theta(x_t) - \theta(x^*) \leq \frac{LR}{\sqrt{T}} .$$

Moreover, going back to (4.1), we get

$$\eta_t (\theta(x_T) - \theta(x^*)) \leq \frac{1}{2} \left( L^2 \eta^2 + \frac{R^2}{T} \right) = \eta LR \frac{\eta}{\sqrt{T}} .$$

\hspace{1cm} \hfill \Box

Remark 51. Remark that, from (4.2), the choice $\eta_t = GR/(L\sqrt{T})$ and $T_0 = T/2$ would give

$$f \left( \frac{\sum_{t=T_0}^T \eta_t x_t}{\sum_{t=T_0}^T \eta_t} \right) - f(x^*) \leq \frac{1}{2} \left( L^2 \frac{\sum_{t=T/2}^T \eta_t^2}{\sum_{t=T/2}^T \eta_t} + \frac{R^2}{\sum_{t=T/2}^T \eta_t} \right) \leq C(G) \frac{RL}{\sqrt{T}} .$$

Hence, it is possible to build an estimator with similar rates of convergence in terms of $R, L$ and $T$ with step sizes independent of $T$.  

4.3 Projected gradient descent

If one wants to minimize a function $\theta$ over a closed convex subset of $\mathbb{R}^d$, it seems natural to build a sequence of estimates taking values in $\mathcal{C}$. In general, gradient descent algorithm may build sequences taking values outside $\mathcal{C}$. To avoid this, one can at each step choose $x_{t+1}$ as the projection of $x_t - \eta_t g_t$ onto $\mathcal{C}$. For this to make sense, let us first show that such a projection exists.

**Theorem 52 (Projection onto closed convex sets).** Let $\mathcal{C}$ denote a closed convex subset of $\mathbb{R}^d$. For any $x \in \mathbb{R}^d$, there exists an element $\pi_\mathcal{C}(x) \in \mathcal{C}$ such that

$$\forall y \in \mathcal{C}, \quad \|x - \pi_\mathcal{C}(x)\| \leq \|x - y\| .$$

Moreover, $\pi_\mathcal{C}(x)$ is the unique element of $\mathcal{C}$ such that

$$\forall y \in \mathcal{C}, \quad \langle \pi_\mathcal{C}(x) - x, \pi_\mathcal{C}(x) - y \rangle \leq 0 .$$

**Proof.** Let $x_0 \in \mathcal{C}$, $r = \|x - x_0\|$ and $B = \{y \in \mathbb{R}^d : \|y - x\| \leq r\}$. The set $\mathcal{C} \cap B$ is compact and, clearly

$$\min_{y \in \mathcal{C}} \|x - y\| = \min_{y \in \mathcal{C} \cap B} \|x - y\| .$$

As the function $\|x - \cdot\|$ is continuous on the compact $\mathcal{C} \cap B$, it achieves its minimum and there exists a minimzer $\pi_\mathcal{C}(x)$ of this function. Moreover, for any $y \in \mathcal{C}$ and $t \in (0, 1)$, $\pi_\mathcal{C}(x) - t(\pi_\mathcal{C}(x) - y) \in \mathcal{C}$, so

$$\|x - \pi_\mathcal{C}(x)\|^2 \leq \|x - \pi_\mathcal{C}(x) + t(\pi_\mathcal{C}(x) - y)\|^2$$

$$= \|x - \pi_\mathcal{C}(x)\|^2 + t^2\|\pi_\mathcal{C}(x) - y\|^2 - 2t \langle \pi_\mathcal{C}(x) - x, \pi_\mathcal{C}(x) - y \rangle .$$

Therefore, for any $t \in (0, 1)$,

$$\langle \pi_\mathcal{C}(x) - x, \pi_\mathcal{C}(x) - y \rangle \leq t\|\pi_\mathcal{C}(x) - y\|^2 .$$

Letting $t \to 0$ shows that

$$\langle \pi_\mathcal{C}(x) - x, \pi_\mathcal{C}(x) - y \rangle \leq 0 .$$

Imagine now that there exists a point $\pi \in \mathcal{C}$ such that

$$\forall y \in \mathcal{C}, \quad \langle \pi - x, \pi - y \rangle \leq 0 .$$

Thus, in particular,

$$\langle \pi_\mathcal{C}(x) - x, \pi_\mathcal{C}(x) - \pi \rangle \leq 0 ,$$

$$\langle x - \pi, \pi_\mathcal{C}(x) - \pi \rangle \leq 0 .$$
4.3. PROJECTED GRADIENT DESCENT

Summing these inequalities shows that
\[ \| \pi_C(x) - \pi \|^2 \leq 0. \]

This shows unicity of \( \pi_C(x) \) and concludes the proof of the theorem. \( \square \)

We can now provide the “projected” gradient descent algorithm.

\begin{algorithm}
\textbf{input} : \( x_1 \in C \): initial point, \( (\eta_t)_t \): step sizes, \( T \): stopping time
\begin{algorithmic}[1]
\State for \( t = 1 \) to \( t = T \), do
\State \hspace{1em} find \( g_t \in \partial \theta(x_t) \),
\State \hspace{1em} \( y_{t+1} = x_t - \eta_t g_t \),
\State \hspace{1em} \( x_{t+1} = \pi_C(y_{t+1}) \).
\end{algorithmic}
\State \textbf{Return} \( x_T = T^{-1} \sum_{i=1}^T x_i \) or \( x_T \in \arg\min_{x \in \{x_1, \ldots, x_T\}} \theta(x) \).
\end{algorithm}

\textbf{Algorithm 2}: Projected gradient descent (PGD).

The following result shows performance of the projected gradient descent algorithm.

\textbf{Theorem 53}. Let \( C \) denote a compact non-empty subset of \( \mathbb{R}^d \) with diameter smaller than \( R \). Assume that \( \theta \) is convex, \( L \)-Lipschitz and denote by \( x^* \in \arg\min_{x \in C} \theta(x) \). If, for any \( t \geq 1 \), \( \eta_t = \eta = R/(L\sqrt{T}) \), then, the outputs of PGD satisfy
\[ \theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}, \quad \theta(x_T) - \theta(x^*) \leq \frac{LR}{\sqrt{T}}. \]

\textbf{Proof}. The proof starts as the one for gradient descent.
\[ \theta(x_t) - \theta(x^*) \leq g_t^T (x_t - x^*) \]
\[ = \frac{1}{\eta} (x_t - y_{t+1})^T (x_t - x^*) \]
\[ = \frac{1}{\eta} \left( \|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|x^* - y_{t+1}\|^2 \right) \quad (4.3) \]

Now,
\[ \|x^* - y_{t+1}\|^2 = \|x^* - x_{t+1}\|^2 + \|x_{t+1} - y_{t+1}\|^2 - 2 \langle x_{t+1} - x^*, x_{t+1} - y_{t+1} \rangle. \]

As \( x^* \in C \) and \( x_{t+1} = \pi_C(y_{t+1}) \), by Theorem 52,
\[ \langle x_{t+1} - x^*, x_{t+1} - y_{t+1} \rangle \leq 0. \]
hence,
\[ \|x^* - y_{t+1}\|^2 \geq \|x^* - x_{t+1}\|^2 + \|x_{t+1} - y_{t+1}\|^2 \geq \|x^* - x_{t+1}\|^2 . \]
Plugging this bound into (4.3) shows that
\[ \theta(x_t) - \theta(x^*) \leq \frac{1}{\eta^2} \left( \|g_t\|^2 + \|x_t - x^*\|^2 - \|x^* - x_{t+1}\|^2 \right) . \]
Summing up yields
\[ \frac{1}{T} \sum_{t=1}^T \theta(x_t) - \theta(x^*) \leq \frac{1}{\eta^2} \left( L^2 + \frac{\|x_1 - x^*\|^2}{T} \right) \leq \frac{1}{\eta^2} \left( L^2 + \frac{R^2}{T} \right) = \frac{RL}{\sqrt{T}} . \]
Then both \( \theta(\mathbf{x}_T) \) (by convexity of \( \theta \)) and \( \theta(\mathbf{x}_T) \) (by definition) are smaller than \( T^{-1} \sum_{t=1}^T \theta(x_t) \), which concludes the proof.

4.4 Examples

4.4.1 SVM

Projected gradient descent is very well adapted to minimize functions over convex sets described by Euclidean geometries. This is the case, typically, of SVM where one has to minimize
\[ \min_{\alpha \in \mathbb{R}^N, \alpha^T \mathbb{K} \alpha \leq r^2} \Theta_N(\alpha), \quad \text{where} \quad \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^N \max(0, 1 - Y_i^{(s)} \alpha^T \mathbb{K} e_i) . \]
We have
\[ \nabla \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^N 1_{\{1 - Y_i^{(s)} \alpha^T \mathbb{K} e_i \geq 0\}} Y_i^{(s)} \mathbb{K} e_i , \]
therefore,
\[ \|\nabla \Theta_N(\alpha)\| \leq \frac{1}{N} \sum_{i=1}^N \|\alpha^T \mathbb{K} e_i\| \leq \frac{r}{N} \sum_{i=1}^N \|\mathbb{K} e_i\| . \]
Now, as, for any \( g \in W \), \( \sup_{x \in \mathcal{X}} |g(x)| \leq \|g\|_W \sqrt{\kappa_\infty} \), and \( \|K(x, \cdot)\|_W^2 = K(x, x) \leq \kappa_\infty \),
\[ \|\mathbb{K} e_i\| = \sqrt{\sum_{j=1}^N K(X_i, X_j)^2} \leq \sqrt{\sum_{j=1}^N \|K(X_i, \cdot)\|_W^2 \kappa_\infty} \leq \kappa_\infty \sqrt{N} . \]
4.4. EXAMPLES

Therefore, \( \Theta_N \) is \( L \)-Lipschitz, with \( L = r k_{\infty} \sqrt{N} \). To apply our general bound, it is necessary to compute the diameter \( R \) of the set \( C = \{ \alpha \in \mathbb{R}^N, \alpha^T K \alpha \leq r^2 \} \). Denote by \( \lambda_{\min}(K) \) the smallest eigenvalue of the positive matrix \( K \). For any \( \alpha \in C \),

\[
    r^2 \geq \alpha^T K \alpha \geq \lambda_{\min}(K) \alpha^T \alpha .
\]

Hence,

\[
    R = 2 \sup_{\alpha \in C} \sqrt{\alpha^T \alpha} \leq \frac{2r}{\sqrt{\lambda_{\min}(K)}} .
\]

By Theorem 53, it follows that

\[
    \Theta_N(f_T) - \Theta_N(\tilde{f}_\varphi) \leq \frac{LR}{\sqrt{T}} = \frac{2r^2 k_{\infty} \sqrt{N}}{\sqrt{T} \lambda_{\min}(K)} .
\]

This error is upper bounded by \( 1/N \) if the number of steps in the PGD algorithm satisfies

\[
    T \geq \frac{4r^4 k_{\infty}^2 N^3}{\lambda_{\min}(K)} .
\]

4.4.2 Boosting

On the other hand, PGD is much less efficient to minimize functions on convex sets described by non-Euclidean geometries. To illustrate this, consider the boosting example where one has to minimize

\[
    \min_{\alpha \in B_1} \Theta_N(\alpha), \quad \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y^{(s)}_i \alpha^T X_i) .
\]

The sub-gradients of \( \Theta_N \) are easy to compute

\[
    g\Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} g\varphi(-Y^{(s)}_i \alpha^T X_i)(-Y^{(s)}_i X_i) .
\]

Recall that \( X_i \) is a vector in \( \mathbb{R}^d \) with entries \( f_j(X_i) \) for some soft classifiers \( f_j \). Thus \( \|X_i\|_\infty \leq 1 \) and therefore \( \|X_i\| \leq \sqrt{d} \). Moreover, as \( \varphi \) is \( L \)-Lipschitz, its sub-gradient \( |g\varphi(x)| \leq L \) at any \( x \in \mathbb{R} \). It follows that

\[
    \|g\Theta_N(\alpha)\| \leq L \sqrt{d} .
\]

Therefore, \( \Theta_N \) is \( L' \)-Lipschitz, with \( L' = L \sqrt{d} \). Moreover, the diameter of \( B_1 \) is \( R = 2 \), so the error made by PGD after \( T \) steps is

\[
    \frac{RL'}{\sqrt{T}} = \frac{2L \sqrt{d}}{\sqrt{T}} .
\]
It is smaller than $1/N$ if $T \geq 4L^2dN^2$. In large dimensional problems where $d \gg N$, this algorithm can be improved substantially.

## 4.5 Mirror descent

Mirror descent algorithms are extensions of projected gradient descent algorithms that can be adapted to non-Euclidean geometries. The first idea of this extension is to bound inner products $a^T b$ using other duality formulae than Cauchy-Schwarz inequality.

**Definition 54.** Let $| \cdot |$ denote a norm on $\mathbb{R}^d$. The dual norm $| \cdot |_*$ is defined by

$$
\forall c \in \mathbb{R}^d, \quad |c|_* = \sup_{x \in \mathbb{R}^d : |x| \leq 1} x^T c .
$$

In particular, for any norm $| \cdot |$ and any $a, b$ in $\mathbb{R}^d$, one has $a^T b \leq |a||b|_*$. The second idea to extend PGD is to introduce potential functions.

**Definition 55 (Potential functions).** A function $\Phi$ defined on a convex domain $C$ is called a potential if there exist a constant $\nu > 0$ and a norm $| \cdot |$ such that $\Phi$ is $\nu$-strongly convex with respect to the norm $| \cdot |$, that is, if it is convex and

$$
\forall x, y \in C, \forall g \in \partial \Phi(x), \quad \Phi(y) \geq \Phi(x) + g^T (y - x) + \frac{\nu}{2} |y - x|^2 .
$$

This definition strengthens the notion of subgradient which only implies that

$$
\forall x, y \in C, \forall g \in \partial \Phi(x), \quad \Phi(y) \geq \Phi(x) + g^T (y - x) .
$$

**Example 1** Let $C = \mathbb{R}^d$, $\Phi(x) = \|x\|^2/2$. Then $\Phi$ is differentiable with gradient $\nabla \Phi(x) = x$. For any $y \in \mathbb{R}^d$,

$$
\Phi(y) = \frac{\|y\|^2}{2} = \frac{\|x\|^2}{2} + x^T (y - x) + \frac{\|y - x\|^2}{2} = \Phi(x) + \nabla \Phi(x)^T (y - x) + \frac{1}{2} \|y - x\|^2 .
$$

Therefore, $\Phi$ is 1-strongly convex with respect to the Euclidean norm over $\mathbb{R}^d$. Therefore, it is a potential and, as will be clear, mirror descent is an extension of PGD.
Example 2  Let $C = \Delta_d$ and let $\Phi(x) = \sum_{i=1}^{d} x_i \log(x_i)$. Then $\nabla \Phi(x) = (\log(x_i) + 1)_{1 \leq i \leq d}$ and, for any $y \in \mathbb{R}^d$,

\[
\Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x) = \sum_{i=1}^{d} y_i \log(y_i) - x_i \log(x_i) - (\log(x_i) - 1)(y_i - x_i)
\]

\[
= \sum_{i=1}^{d} y_i \log \left( \frac{y_i}{x_i} \right) + y_i - x_i
\]

\[
= \sum_{i=1}^{d} y_i \log \left( \frac{y_i}{x_i} \right).
\]

By Pinsker’s inequality, see Lemma 36,

\[
\Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x) \geq \frac{1}{2} \left( \sum_{i=1}^{d} |x_i - y_i| \right)^2 = \frac{\|x - y\|^2}{2}.
\]

Therefore, $\Phi$ is 1-strongly convex with respect to $\| \cdot \|_1$ over $\Delta_d$.

A key step to analyse gradient descent and PGD is the formula

\[
a^T b = \frac{1}{2} \left( \|a\|^2 + \|b\|^2 - \|a - b\|^2 \right).
\]

This formula can be extended to more general potential functions using Bregman divergences.

Definition 56 (Bregman divergence). Let $\Phi$ denote a convex function over a convex domain $C$. Then, for any $x, y$ in $C$ the Bregman divergence of $\Phi$ from $y$ to $x$ is defined by

\[
D_\Phi(y, x) = \Phi(y) - \Phi(x) - \nabla \Phi(x)^T (y - x).
\]

$D_\Phi(y, x)$ is therefore the error made when approximating $\Phi(y)$ by the linear approximation $\Phi(x) + \nabla \Phi(x)^T (y - x)$. When $\Phi$ is a potential, its Bregman divergence acts as the square of the norm $\| \cdot \|$. The key to extend formula (4.6) is the following proposition.

Proposition 57. If $\Phi$ is convex on a convex domain $C$ and $x, y, z$ lie in $C$, then

\[
D_\Phi(y, x) + D_\Phi(x, y) = (\nabla \Phi(y) - \nabla \Phi(x))^T (y - x),
\]

\[
(\nabla \Phi(x) - \nabla \Phi(y))^T (x - z) = D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y).
\]
Proof. The proof is completely elementary. The first equality is immediate and, for the second one,
\[
D_\Phi(x, y) + D_\Phi(z, x) - D_\Phi(z, y) = \Phi(x) - \Phi(y) - \nabla \Phi(y)^T(x - y) \\
+ \Phi(z) - \Phi(x) - \nabla \Phi(x)^T(z - x) \\
- \nabla \Phi(z) + \Phi(y) + \nabla \Phi(y)^T(z - y) \\
= \nabla \Phi(y)^T(z - y - x + y) - \nabla \Phi(x)^T(z - x) \\
= (\nabla \Phi(y) - \nabla \Phi(x))^T(z - x) .
\]

As for projected gradient descent, we also need to project according to the potential $\Phi$.

\textbf{Definition 58 (Bregman projection).} Let $\Phi$ denote a convex function defined on a closed convex domain $C$ and let $C$ denote a convex subset of $C$. The Bregman projection of any $x \in C$ on $C$ with respect to the function $\Phi$ is
\[
\pi_C^\Phi(x) \in \arg\min_{y \in C} D_\Phi(x, y) .
\]

\textbf{Example 1} The Bregman divergence with respect to $\Phi = \| \cdot \|^2 / 2$ is given by (see Eq 4.4)
\[
D_\Phi(y, x) = \frac{\| x - y \|^2}{2} .
\]

Therefore, Bregman projection with respect to $\Phi$ is the usual projection onto convex sets shown in Theorem 52.

\textbf{Example 2} Let us consider $C = (\mathbb{R}_+)^d$, $C = \Delta_d \subset C$ and $\Phi(x) = \sum_{i=1}^d x_i \log(x_i)$. In this example, the Bregman divergence is given, see (1.3), by
\[
D_\Phi(y, x) = \sum_{i=1}^d y_i \log \left( \frac{y_i}{x_i} \right) + y_i - x_i .
\]

The Bregman projection $x$ of $y \in C$ is a mimizer of $D_\Phi(y, \cdot)$ over $\Delta_d$. Define the Lagragian
\[
\mathcal{L}_\lambda(x) = D_\Phi(y, x) + \lambda \left( \sum_{i=1}^d x_i - 1 \right) .
\]

It holds
\[
\forall i \in \{1, \ldots, d\}, \quad \frac{\partial \mathcal{L}_\lambda}{\partial x_i}(x) = - \frac{y_i}{x_i} + 1 + \lambda, \quad \frac{\partial \mathcal{L}_\lambda}{\partial \lambda}(x) = \sum_{i=1}^d x_i - 1 .
\]
Solving the equations \(\frac{\partial \mathcal{L}}{\partial x_i}(\mathbf{x}) = 0\), \(\frac{\partial \mathcal{L}}{\partial \lambda}(\mathbf{x}) = 0\) yields, for all \(i\), \(x_i = y_i/(\lambda - 1)\) and \((\lambda - 1) = \sum_{i=1}^d y_i = \|\mathbf{y}\|_1\), so finally

\[
\pi_{\Delta_d}(\mathbf{y}) = \mathbf{x} = \frac{\mathbf{y}}{\|\mathbf{y}\|_1}.
\]

As for the usual projections onto convex sets, Bregman projections have a nice characterization in terms of inner products.

**Proposition 59.** For any \(z \in C\) and any \(y \in C\),

\[
(\nabla \Phi(\pi_C^\phi(y)) - \nabla \Phi(y))^T (\pi_C^\phi(y) - z) \leq 0.
\]

In addition,

\[
D_\phi(z, \pi_C^\phi(y)) \leq D_\phi(z, y).
\]

**Proof.** For any \(t \in [0, 1]\), let \(h(t) = D_\phi(\pi_C^\phi(y) + t(x - \pi_C^\phi(y)), y)\). As \(\pi_C^\phi(y) + t(x - \pi_C^\phi(y)) \in C\) by convexity, \(h\) is minimal at \(t = 0\), so

\[
0 \leq h'(0) = \nabla_x D_\phi(x, y)|_{x = \pi_C^\phi(y)}(z - \pi_C^\phi(y)).
\]

Now, by construction of \(D_\phi\),

\[
\nabla_x D_\phi(x, y)|_{x = \pi_C^\phi(y)} = \nabla \Phi(\pi_C^\phi(y)) - \nabla \Phi(y).
\]

Hence,

\[
(\nabla \Phi(\pi_C^\phi(y)) - \nabla \Phi(y))^T (z - \pi_C^\phi(y)) \geq 0.
\]

This proves the first item. For the second, it follows from Proposition 57 and the first item that

\[
D_\phi(\pi_C^\phi(y), y) + D_\phi(z, \pi_C^\phi(y)) - D_\phi(z, y) \leq 0.
\]

As \(D_\phi(\pi_C^\phi(y), y) \geq 0\), this implies that

\[
D_\phi(z, \pi_C^\phi(y)) \leq D_\phi(z, y).
\]

This concludes the proof of the proposition. \(\square\)

We are now ready to give the mirror descent algorithm
CHAPTER 4. CONVEX OPTIMIZATION

Algorithm 3: Mirror descent (MD).

1. \textbf{input} \: x_1 \in \arg\min_{x \in C} \Phi(x): \text{initial point,} \: (\eta_t)_t: \text{step sizes,} \: T: \text{stopping time}

2. \textbf{for} \: t = 1 \text{ to } t = T, \text{ do}

3. \quad \text{find } g_t \in \partial \theta(x_t),

4. \quad \nabla \Phi(y_{t+1}) = \nabla \Phi(x_t) - \eta_t g_t,

5. \quad x_{t+1} = \pi_C(y_{t+1}).

6. \textbf{end}

\textbf{Return} \: \bar{x}_T = T^{-1} \sum_{t=1}^{T} x_t \text{ or } \bar{x}_T \in \arg\min_{x \in \{x_1, \ldots, x_T\}} \theta(x).

Compared with gradient descent, mirror descent uses the gradients of the potential \( \Phi \) to map elements of \( \mathbb{R}^d \) endowed with the norm \( \| \cdot \| \) (called primal) into its dual which is homeomorphic to \( \mathbb{R}^d \) endowed with the dual norm \( \| \cdot \|_* \).

Then, mirror descent uses a gradient descent step in the dual space and maps back the images into the primal space. The term mirror originates from this mapping into the dual space. One can analyse MD algorithm almost exactly as PGD.

\textbf{Theorem 60.} Assume that \( \theta \) is convex on a (closed convex) domain \( C \) and that it is \( L \)-Lipschitz with respect to the norm \( \| \cdot \| \), that is \( |g_t|_* \leq L \) and any \( g_t \in \partial \theta(x) \) and any \( x \in C \). Let \( C \) denote a closed convex subset of \( C \). Assume that \( \Phi \) is \( \nu \)-strongly convex on \( C \) with respect to \( \| \cdot \| \). Let \( x^* \in \arg\min_{x \in C} \theta(x) \) and

\[ R^2 = \sup_{x \in C} \Phi(x) - \inf_{x \in C} \Phi(x). \]

Then, the outputs of MD algorithm with \( \eta_t = (R/L)\sqrt{2\nu/T} \) satisfy

\[ \theta(\bar{x}_T) - \theta(x^*) \leq RL \sqrt{\frac{2}{\nu T}}, \quad \theta(\bar{x}_T) - \theta(x^*) \leq RL \sqrt{\frac{2}{\nu T}}. \]

\textbf{Proof.} Let \( x_o \in C \cap C \), by definition of the MD update,

\[ \theta(x_t) - \theta(x_o) \leq g_t^T (x_t - x_o) = \frac{1}{\eta} (\nabla \Phi(x_t) - \nabla \Phi(y_{t+1}))^T(x_t - x_o). \]

By Proposition 57, it follows that

\[ \theta(x_t) - \theta(x_o) \leq \frac{1}{\eta} \left(D_{\Phi}(x_t, y_{t+1}) + D_{\Phi}(x_o, x_t) - D_{\Phi}(x_o, y_{t+1})\right). \]

Now we invoque the following result.
Proposition 61. Using updates of MD, if $\Phi$ is $\nu$-strongly convex with respect to $| \cdot |$, then

$$D_\Phi(x_t, y_{t+1}) \leq \frac{\eta_g^2}{2\nu}.$$ 

Proof of Proposition 61. By Proposition 57,

$$D_\Phi(x_t, y_{t+1}) = -D_\Phi(y_{t+1}, x_t) + (\nabla \Phi(x_t) - \nabla \Phi(y_{t+1})^T(x_t - y_{t+1}).$$

By strong convexity, it holds therefore

$$D_\Phi(x_t, y_{t+1}) = -\frac{\nu}{2} |y_{t+1} - x_t|^2 + (\nabla \Phi(x_t) - \nabla \Phi(y_{t+1})^T(x_t - y_{t+1}).$$

Using the MD updates and the definition of the dual norm, one gets

$$D_\Phi(x_t, y_{t+1}) \leq \eta g_t^T(x_t - y_{t+1}) - \frac{\nu}{2} |y_{t+1} - x_t|^2$$

$$\leq \eta |g_t|^\ast |x_t - y_{t+1}| - \frac{\nu}{2} |y_{t+1} - x_t|^2.$$ 

The proof terminates by the inequality $\sup_{x \in \mathbb{R}} \{ax - bx^2/2\} = a^2/(2b)$, which holds for any $a \in \mathbb{R}$ and $b > 0$.

By Proposition 61,

$$D_\Phi(x_t, y_{t+1}) \leq \frac{\eta^2 |g_t|^2}{2\nu}.$$ 

As $\theta$ is $L$-Lipschitz, this implies that

$$D_\Phi(x_t, y_{t+1}) \leq \frac{\eta^2 L^2}{2\nu}.$$ 

Moreover, by Proposition 54

$$D_\Phi(x_o, x_{t+1}) \leq D_\Phi(x_o, y_{t+1}).$$ 

Hence,

$$\theta(x_t) - \theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{D_\Phi(x_o, x_t) - D_\Phi(x_o, x_{t+1})}{\eta}.$$ 

Summing up

$$\frac{1}{T} \sum_{t=1}^{T} \theta(x_t) - \theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{D_\Phi(x_o, x_1)}{\eta T}.$$
As $x_1$ is a minimizer of $\Phi$ over $C$,
\[ D_\Phi(x_o, x_1) = \Phi(x_o) - \Phi(x_1) - \nabla \Phi(x_1)^T (x_o - x_1) \leq \Phi(x_o) - \Phi(x_1) \leq R^2. \]
Thus,
\[ \frac{1}{T} \sum_{t=1}^{T} \theta(x_t) - \theta(x_o) \leq \frac{\eta L^2}{2\nu} + \frac{R^2}{\eta T} \leq LR\sqrt{\frac{2}{\nu T}}. \]
The proof concludes as usual that
\[ \theta(x_T) - \theta(x_o) \leq LR\sqrt{\frac{2}{\nu T}}, \quad \theta(x_T) - \theta(x_o) \leq LR\sqrt{\frac{2}{\nu T}}. \]
Finally, one can let $x_o \to x^*$ to obtain the theorem.

To show the interest of MD, let us go back to the boosting example. Take $\Phi(x) = \sum_{i=1}^{d} x_i \log(x_i)$, so the MD update is equivalent to
\[ \forall i \in \{1, \ldots, d\}, \quad \log(y_{t+1,i}) = \log(x_{t,i}) - \eta g_{t,i}, \quad \text{i.e.} \quad y_{t+1,i} = x_{t,i} e^{-\eta g_{t,i}}, \]
and $x_{t+1} = y_{t+1}/\|y_{t+1}\|_1$. Recall that the function to minimize is
\[ \Theta_N(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y^{(s)}_i \alpha^T X_i), \]
where the vectors $X_i$ are defined thanks to a collection of preliminary soft classifiers $f_1, \ldots, f_d$ by
\[ X_i = \begin{bmatrix} f_1(X_i) \\ \vdots \\ f_d(X_i) \end{bmatrix}. \]
The sub-gradients of this objective function can be computed:
\[ g_{\Theta_N}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} \varphi(-Y^{(s)}_i \alpha^T X_i)(-Y^{(s)}_i X_i). \]
Now $\| \cdot \| = \| \cdot \|_1$ has dual norm $\| \cdot \|_* = \| \cdot \|_\infty$. As $\varphi$ is $L$-Lipschitz, its subgradients $|g_\varphi(x)| \leq L$ at any point $x$. As $Y^{(s)}_i \in [-1, 1]$ and $\|X_i\|_\infty \leq 1$, it holds that
\[ \|g_{\Theta_N}(\alpha)\|_\infty \leq L. \]
As this holds for any $\alpha \in \Delta_d$, this shows that $\Theta_N$ is $L$-Lipschitz with respect to $\| \cdot \|_1$. Moreover, for any $x \in \Delta_d$, $\Phi(x) \leq 0$ and
\[ \Phi(x) + \log d = \sum_{i=1}^{d} x_i \log(x_i/(1/d)) \geq 0, \]
4.6. STOCHASTIC OPTIMIZATION

so

\[ \Phi(x) \geq -\log d . \]

Hence, \( R^2 = \log d \). Theorem 63 shows then that MDA has, after \( T \) steps, an error upper bounded by

\[ L \sqrt{\frac{2 \log d}{T}} . \]

This error is upper bounded by \( 1/N \) as soon as \( T \geq 2L^2N^2 \log d \). Recall that PGD on this example achieved this error after \( 4L^2dN^2 \) steps. When the dimension \( d \) is large, the improvement brought by MD over PGD is substantial.

4.6. Stochastic optimization

In this section, consider functions \( \theta: \mathcal{C} \times \mathcal{Z} \to \mathbb{R} \) such that \( \mathcal{C} \) is a convex subset of \( \mathbb{R}^d \) and, given a distribution \( P \) on \( \mathcal{Z} \), for \( P \)-almost \( z \in \mathcal{Z} \), the function \( \theta(\cdot, z) \) is convex. Assume also that, for any \( x \in \mathcal{C} \), \( E[|\theta(x, Z)|] < \infty \), so the convexity assumption implies that the function \( \Theta = E[\theta(\cdot, Z)] \) is convex. The goal of stochastic optimization is to approximate

\[ \min_{x \in \mathcal{C}} \Theta(x) \]

using an i.i.d. sample \( Z_1, \ldots, Z_T \) with distribution \( P \). Remark that the difference with convex optimization is that the objective function \( \Theta \) is unknown and we only have access to noisy versions of it (and its gradients), therefore, one cannot define the estimators

\[ x_T \in \arg \min_{x \in \{x_1, \ldots, x_T\}} \Theta(x) \]

that we used in the convex optimization algorithms. On the other hand, by focusing on the actual expected value of the loss, the performance of the final estimator won’t rely on those of the ERM.

4.6.1 Stochastic gradient descent

Denote by \( \partial \theta(x, Z) \) the sub-differential of \( \theta(\cdot, Z) \) at point \( x \in \mathbb{R} \). The stochastic gradient descent algorithm is described below
CHAPTER 4. CONVEX OPTIMIZATION

input: $x_1 \in C$: initial point, $(\eta_t)$: step sizes, $T$: stopping time
$Z_1, \ldots, Z_T$ i.i.d. random variables

1. for $t = 1$ to $t = T$, do
   2. find $\tilde{g}_t \in \partial \theta(x_t, Z_t)$,
   3. $y_{t+1} = x_t - \eta_t \tilde{g}_t$,
   4. $x_{t+1} = \pi_C(y_{t+1})$.
5. end
6. Return $\bar{x}_T = T^{-1} \sum_{t=1}^T x_t$.

Algorithm 4: Stochastic gradient descent (SGD).

An important difference between gradient descent and stochastic gradient
descent is that all gradients $\partial \theta(x_t, Z_t)$ for $t = 1, \ldots, T$ must be available
for gradient descent while only one of these gradients is used for stochastic
gradient descent. This is particularly interesting for huge datasets as these
gradients have to be computed in different servers and may be stored in
different places. In this case, requesting all gradients can be both demanding
on memory resources and produce non-reliable estimators if a server is very
slow or crash down. In these situations, stochastic gradient is much more
efficient.

Theorem 62. Let $C$ denote a closed, convex subset of $\mathbb{R}^d$ with diameter
smaller than $R$. Let $x^* \in \text{argmin}_{x \in C} \Theta(x)$. Assume that, for any $x \in C$,
y any random subgradient $\tilde{g}_t \in \partial \theta(x, Z)$ satisfies $E[\|\tilde{g}_t\|^2] \leq L^2$. If, for any
$t = 1, \ldots, T$, $\eta_t = \eta = R/(L\sqrt{T})$, it holds

$$E[\Theta(x_T)] - \Theta(x^*) \leq \frac{RL}{\sqrt{T}}.$$ 

Proof. Let $\mathcal{F}_t$ denote the $\sigma$-algebra generated by $Z_1, \ldots, Z_{t-1}$ so $x_t$ is $\mathcal{F}_t$-
measurable. Let $\tilde{g}_t \in \partial \theta(x_t, Z_t)$ so, by independence of $Z_t$ and $x_t$,

$$g_t = E[\tilde{g}_t|\mathcal{F}_t] \in \partial \Theta(x_t).$$

Then,

$$\Theta(x_t) - \Theta(x^*) \leq g_t^T (x_t - x^*)$$
$$= E[\tilde{g}_t^T (x_t - x^*)|\mathcal{F}_t]$$
$$= \frac{1}{\eta} E[(x_t - y_{t+1})^T (x_t - x^*)|\mathcal{F}_t]$$
$$= \frac{1}{2\eta} E[(\|x_t - y_{t+1}\|^2 + \|x_t - x^*\|^2 - \|y_{t+1} - x^*\|^2)|\mathcal{F}_t]$$
$$\leq \frac{1}{2\eta} E[(\eta^2 \|\tilde{g}_t\|^2 + \|x_t - x^*\|^2 - \|x_{t+1} - x^*\|^2)|\mathcal{F}_t].$$
4.6. STOCHASTIC OPTIMIZATION

Taking expectations,

\[ E[\Theta(x_t)] - \Theta(x^*) \leq \frac{\eta L^2}{2} + \frac{E[\|x_t - x^*\|^2] - E[\|x_{t+1} - x^*\|^2]}{2}. \]

We conclude as before by summing up, bounding \( \|x_1 - x^*\|^2 \leq R^2 \), and using the convexity of \( \Theta \). \( \square \)

4.6.2 Stochastic mirror descent

Stochastic gradient descent suffers the same issues as gradient descent when \( \mathcal{C} \) is a convex adapted to \( \ell_1 \)-geometry rather than Euclidean geometry. Fortunately, a stochastic version of general mirror descent algorithms can easily be designed to bypass this issue.

```
input : \( x_1 \in \text{argmin}_{x \in \mathcal{C}} \Phi(x) \): initial point, \((\eta_t)_t\): step sizes,
\( Z_1, \ldots, Z_T \): dataset
1 for \( t = 1 \) to \( t = T \), do
  2 find \( \tilde{g}_t \in \partial \theta(x_t, Z_t) \),
  3 \( \nabla \Phi(y_{t+1}) = \nabla \Phi(x_t) - \eta_t \tilde{g}_t \),
  4 \( x_{t+1} = \pi_{\mathcal{C}}^\Phi(y_{t+1}) \).
5 end
6 Return \( \bar{x}_T = T^{-1} \sum_{t=1}^T x_t \) or \( x_T \in \text{argmin}_{x \in \{x_1, \ldots, x_T\}} \theta(x) \).
```

Algorithm 5: Stochastic mirror descent (SMD).

Mixing the proofs of SGD and MD yields the following result.

**Theorem 63.** Assume that \( \theta(\cdot, Z) \) is \( P \)-almost surely convex on a domain \( \mathcal{C} \) and that, for any \( x \in \mathcal{C} \), any random subgradient \( \tilde{g}_t \in \partial \theta(x, Z) \) satisfies \( E[\|\tilde{g}_t\|_2^2] \leq L^2 \). Let \( \mathcal{C} \) be a closed convex subset \( \mathcal{C} \). Assume that \( \Phi \) is \( \nu \)-strongly convex on \( \mathcal{C} \) with respect to \( \| \cdot \| \). Let \( x^* \in \text{argmin}_{x \in \mathcal{C}} \theta(x) \).

\[
R^2 = \sup_{x \in \mathcal{C}} \Phi(x) - \inf_{x \in \mathcal{C}} \Phi(x).
\]

Then, the outputs of SMD with \( \eta_t = (R/L)\sqrt{2\nu/T} \) satisfy

\[
\Theta(\bar{x}_T) - \Theta(x^*) \leq RL \sqrt{\frac{2}{\nu T}}.
\]

**Proof.** The proof starts as for SGD and terminates as for MD. Let \( \mathcal{F}_t \) denote the \( \sigma \)-algebra generated by \( Z_1, \ldots, Z_{t-1} \) so \( x_t \) is \( \mathcal{F}_t \)-measurable. Let \( \tilde{g}_t \in \partial \theta(x_t, Z_t) \) so, by independence of \( Z_t \) and \( x_t \),

\[
g_t = E[\tilde{g}_t | \mathcal{F}_t] \in \partial \Theta(x_t).
\]
Then, for any $x_o \in C$,

$$\Theta(x_t) - \Theta(x_o) \leq g_t^T(x_t - x_o)$$

$$= E[\lambda_t^T(x_t - x_o)]$$

$$= \frac{1}{\eta} E[(\nabla \Phi(x_t) - \nabla \Phi(y_{t+1}))^T(x_t - x_o)]$$

$$\leq \frac{1}{\eta} E[D_\Phi(x_t, y_{t+1}) + D_\Phi(x_o, x_t) - D_\Phi(x_o, y_{t+1})]$$

$$\leq \frac{1}{\eta} E \left[ \frac{\eta^2 |g_t|^2}{2\nu} + D_\Phi(x_o, x_t) - D_\Phi(x_o, x_{t+1}) \right].$$

Taking expectations and summing yields

$$E \left[ \frac{1}{T} \sum_{t=1}^T \Theta(x_t) - \Theta(x_o) \right] \leq \frac{\eta L^2}{2\nu} + \frac{D_\Phi(x_o, x_1)}{\eta T}.$$

As $x_1$ is a minimizer of $\Phi$, $D_\Phi(x_o, x_1) \leq \Phi(x_o) - \Phi(x_1) \leq R^2$, so, by convexity of $\Theta$,

$$E \left[ \Theta(x_t) - \Theta(x_o) \right] \leq \frac{\eta L^2}{2\nu} + \frac{R^2}{\eta T} = \frac{2LR}{\sqrt{2\nu T}}.$$

Conclude by letting $x_o \to x^*$. \qed